TITULNÍ LIST PERIODICKÉ ZPRÁVY 2007 PROJEKTU 2C06009

Ministerstvo školství, mládeže a tělovýchovy

2C06009

PROSTŘEDKY TVORBY KOMPLEXNÍ BÁZE ZNALOSTÍ PRO KOMUNIKACI SE SÉMANTICKÝM WEBEM V PŘIROZENÉM JAZYCE

řešitel - **doc. Ing. Karel Ježek, CSc.**

.....

(podpis)

za příjemce - koordinátor - **Západočeská univerzita v Plzni** (IČ: 49777513)

rektor

Doc. Ing. Josef Průša, CSc.

.....

(podpis, razítko)

Verze zprávy: **1** Zpracováno dne:

2. SKUTEČNOST ZA UPLYNULÉ OBDOBÍ - 2007

2.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

2.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce - koordinátor
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 8/
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

2.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Albrecht Štěpán Ing.** 810520/2061 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 496 377 632 402 albrs@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Brada Přemysl Ing. PhD. MSc.** 7007012111 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632435 brada@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Češka Zdeněk Ing** 8207311244 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 zceska@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	31.25

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Fiala Dalibor Ing.** 8003235845 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 dalfa@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Hanks Patrick Ph.D.** 400324 GB
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	hanks@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Horák Aleš Ph.D.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Hynek Jiří ing. PhD** 720506/2029 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632455 hynekj@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	40

Celé jméno, RČ	**Král Pavel ing. PhD.** 760317/2049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pavelka Tomáš Ing.** 7909182083 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Ptáčková Helena** 705914/2079 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	5

Celé jméno, RČ	**Rychlý Pavel Mgr. Ph.D.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Sojka Petr RNDr. Ph.D.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Steinberger Josef Ing.** 7909182127 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 479 jstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Tesař Roman Ing.** 7909302379 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 romant@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Toman Michal Ing.** 8007042054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 mtoman@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Zíma Martin Ing. PhD.** 7405042073 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632431 zima@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

2.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2007

Pč.	Typ	Popis
1	změny v projektovém týmu a řešitelských týmech	V průběhu roku čtyři členové řešitelského kolektivu ZČU podali své disertační práce, jejichž náplň byla součástí projektu. Tři již své práce úspěšně obhájili. Dva z nich, Josef Steinberger a Pavel Král, pokračují v práci na projektu jako kmenoví pracovníci katedry. Třetí, D. Fiala, odešel do zaměstnání mimo ZČU. Do projektu byli nově zapojeni Z. Češka, věnující se problematice detekce plagiátů a podobnosti dokumentů, M. Zíma a P. Brada pracující na problematice porozumění textu a extrakci znalostí.

2.2. ČASOVÝ POSTUP PRACÍ

Komentář k metodice a časovému postupu prací a průběhu aktivit za uplynulé období

Metodika a časový rozvrh postupu prací byly dodrženy. Během řešení se však vyskytla potřeba doplnění dalších dílčích úkolů, neboť některé dílčí činnosti, jejichž provedení bylo původně plánovano v rámci jiných činností, se ukázaly natolik rozsáhlé, že byly do plánu řešitelských prací a také do předkládané zprávy doplněny jako samostatné. Jinak se plán řešitelských prací nezměnil a byl plně dodržen.

2.2.0. PŘEHLED DÍLČÍCH CÍLŮ SCHVÁLENÉ- SKUTEČNOST 2007

	Číslo	Dílčí cíl	Datum plnění
	1	Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím.	- 31.12.2007
	2	Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka.	- 31.12.2008
	3	Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce.	- 31.12.2009
	4	Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí.	- 31.12.2010

2.2.1. AKTIVITY USKUTEČNĚNÉ v roce 2007

Číslo aktivity

01

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Pořizování korpusu LAC-SS

Zahájení aktivity

2.12.2006

Ukončení aktivity

Popis aktivity

Pořízení a transkripce nahrávek spontánní řeči do korpusu LAC-SS (LICS Audio Corpus – Spontaneous Speech). Postupovalo se podle metodologie vytvořené v rámci aktivity 2006-01 "Příprava pořizování korpusu LAC-SS2006". Korpus LAC-SS (Spontaneous Speech) je určen k trénování recognizeru LASER pro rozpoznávání spontánní plynulé řeči. Obsahuje pouze záznamy spontánní, nepřipravené a nečtené řeči. Po první fázi řešení projektu obsahuje korpus LAC-SS celkem 741 minut (12h21m45s) záznamu spontánní řeči ve 24 nahrávkách vysoké kvality, k nimž byla pořízena manuální transkripce. 11 nahrávek pochází z přednášek pracovníků Katedry informatiky a výpočetní techniky - jedná se tedy o mluvčí s rozsáhlou praxí v oblasti veřejných projevů. Naopak 13 nahrávek pochází ze studentských seminářů - mluvčí (studenti) ve většině případů nejsou zvyklí pronášet veřejné projevy, a proto také záznamy obsahují celou řadu neřečových zvuků a projevů, např. nervozity, což je v tomto případě ale žádoucí, protože díky tomu je korpus dostatečně foneticky rozmanitý a bohatý. Nahrávky byly pořízeny elektretovým klopovým mikrofonem a zaznamenány minidiskovým rekordérem Sony MZ-RH1 ve dvou akusticky různých prostředích: v malé posluchárně a ve velké posluchárně. Díky speciální směrové a frekvenční charakteristice mikrofonu však akustika prostředí nehraje v záznamech podstatnou roli. Z pohledu výpočetní lingvistiky korpus obsahuje celkem 40866 lexikálních atomů, z toho 33795 slov a 7071 neřečových zvuků (non-speech sounds), např. odkašlávání, popotahování, kýchání, mlaskání, apod. Celkový počet různých slov v korpusu je 6894. Manuální transkripci záznamů provedly (a nadále provádějí) tři studentky Filosofické fakulty Západočeské univerzity, které byly vybrány podle výsledků zkušební práce a na základě splněného testu předpokladů. Vybrané studentky (dále jen transkriberky) následně prošly rozsáhlým školením v oblasti české fonetiky a fonologie, transkripce a v nezbytném rozsahu také teorie rozpoznávání řeči, a poté výcvikem v práci s transkripčním softwarem. Pro potřeby transkriberek byl vytvořen tzv. transkripční manuál, který přesně popisuje správný postup práce s transkripčním softwarem Transcriber 1.5.1, obsahuje úplný přehled transkripčních značek a příkladů jejich použití a také návod, jak řešit některé sporné případy přepisu. Další doškolování a průběžné konzultace transkriberek se členy řešitelského týmu se jednoznačně pozitivně projevily na nárůstu rychlosti transkripce, která je v současné době dvojnásobná oproti začátku projektu. Nyní dokáže jedna transkriberka přepsat za hodinu asi 6 minut záznamu. Transkripce je ortograficko-fonetická, tj. základem je ortografický přepis, doplněný speciálními značkami pro neřečové zvuky. V případě, že ortografie není známa nebo v záznamu je zcela prokazatelně slyšitelný jiný tvar slova či zvuku, než jaký je českou ortografií akceptovatelný, je přepis uveden v závorkách foneticky. Podrobnější statistika získaného materiálu je k dispozici v článku [K. Ekštein: On Building of Czech Spontaneous Speech Corpus, SPECOM 2007]. V rámci úlohy pořizování korpusu byla také zahájena práce na sběru záznamů pozaďových ruchů a šumů. Korpus ruchů a šumů je nazván LAC-Noise. Nahrávky z korpusu LAC-Noise budou použity k trénování recognizeru LASER v obtížných příjmových podmínkách. Předpokládaný mód trénování je použití těchto záznamů jako aditivního šumu, který bude smísen s relativně čistými nahrávkami z korpusů LAC. Tímto způsobem je možné získat značné množství hodnotného trénovacího materiálu pouze za cenu strojové práce (smíchání stopy čistého záznamu řeči a pozaďového šumu na zvolené úrovni SNR). Nahrávky byly pořízeny v terénu vysoce citlivým všesměrovým elektretovým mikrofonem Sennheiser MKE 2 a zaznamenány buď přímo zvukovou kartou notebooku nebo minidiskovým rekordérem. Zatím byly pořízeny nahrávky těchto ruchů a šumů: (i) vlak 01 - hluk v kupé rychlíku ČD jedoucím cestovní rychlostí na trati Plzeň - Praha; (ii) autobus 01 - hluk

v luxusním autobusu jedoucím rychlostí 100 km/h po dálnici; (iii) diesel 01 - hluk v osobním automobilu s atmosferickým vznětovým motorem, který se pohyboval v městském provozu; (iv) křižovatka 01 - hluk na frekventované křižovatce (ulic Klatovská a Americká) v centru Plzně v době dopravní špičky; (v) nádražní hala 01 - hluk v nádražní hale plzeňského hlavního nádraží v době dopravní špičky.

Skutečné Indikátory dosažení - výsledky aktivity

Nahrávky a jejich ortografická transkripce.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Nahrávky budou použity k trénování akustických modelů rozpoznávače. Ověřením bude mj. i výsledná úspěšnost rozpoznávání.

Číslo aktivity

02

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Integrace komponent systému LASER

Zahájení aktivity

1.2.2007

Ukončení aktivity

31.12.2007

Popis aktivity

V předchozí verzi byly jednotlivé komponenty rozpoznávače LASER (viz http://liks.fav.zcu.cz/mediawiki/index.php/LASER) implementovány jako programy spustitelné z příkazové řádky, jejichž vstupem a výstupem je datový soubor. Tento způsob implementace je vhodný pro testování a výzkum, nikoli však pro real time aplikaci. Využití souborů jako prostředku pro komunikaci mezi moduly značně prodlužuje dobu odezvy systému, protože 1. čtení souborů z disku je pomalé, 2. ostatní komponenty musejí čekat, než skončí nahrávání (nahrávání není výpočetně náročné a zbylý procesorový čas je možné využít). Na začátku roku 2007 proběhly pokusy o implementaci některých komponent rozpoznávače v jazyce Java. Tyto pokusy se ukázaly jako úspěšné a brzy vznikla první verze rozpoznávače, pojmenovaného JLASER, která obsahovala veškerou funkcionalitu předchozího systému. Veškeré testy a implementace nových prvků jsou nyní prováděny s rozpoznávacím systémem JLASER. Výsledný produkt je snadno integrovatelný do programů napsaných v jazyce Java. Rozpoznávač podporuje dva typy akustických modelů, konkrétně umělé neuronové sítě a směsi Gaussových funkcí (trénované SW balíkem HTK). Kromě vlastního rozpoznávání systém umožňuje automatické značkování a segmentaci trénovacích dat pro neuronové sítě, vytváření rozpoznávacích grafů a automatickou ortograficko-fonetickou konverzi pro češtinu.

Skutečné Indikátory dosažení - výsledky aktivity

Zdrojové kódy programu (JLASER v. 1.1) dostupné na stránkách
http://liks.fav.zcu.cz/mediawiki/index.php/JLASER. Program byl uveřejněn pod licencí GPL v.2.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Měření rychlosti zpracování – porovnání staré a nové verze a porovnání se softwarem HTK bylo uvedeno v článku Pavelka, T., Ekštein, K.: JLASER: An Automatic Speech Recognizer Written in Java, Proc. of the XII. International Conference Speech and Computer (SPECOM'2007), Moscow, Russia, 2007.

Výsledky testů různých akustických modelů provedených s rozpoznávačem JLASER jsou uvedeny v článcích:

• Pavelka, T., Hejtmánek, J.: Context Dependency in Neural Network Based Acoustic Models, Proc. of PhD Workshop 2007, Balatonfüred, Maďarsko, 2007

• Hejtmánek, J., Pavelka, T.: Use of context-dependent units in Czech speech, Proc. of PhD Workshop 2007, Balatonfüred, Maďarsko, 2007

Číslo aktivity

03

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Detektor ticha a řeči (2)

Zahájení aktivity

2.1.2007

Ukončení aktivity

31.7.2007

Popis aktivity

Navazovala na aktivitu z roku 2006 a zahrnuje implementaci rozpoznávacích algoritmů a natrénování modelů. V dialogových systémech při real-time zpracování řečového signálu jsou mezi jednotlivými dotazy a odpověďmi místa, kde řečník nemluví. V těchto místech je zbytečné rozpoznávat, jestli daný zvuk přísluší některému z fonémů nebo tichu. Je výpočetně výhodnější rozpoznávat, jestli přísluší řeči nebo tichu. Jakmile detektor ticha a řeči identifikuje řeč, předá řízení rozpoznávači JLASER. Reálný signál ale obsahuje kromě mluvy řečníka také různé rušivé zvuky (šramot, mluva na pozadí), které nesmějí být identifikovány jako řeč, pokud v tu chvíli řečník nemluví. Cílem aktivity bylo natrénovat modely pro detekci ticha vs. řeči s využitím Gaussovských mixtur a skrytých markovských modelů a dále pak umělé neuronové sítě založené na vícevrstvém perceptronu. Dalším dílčím cílem bylo zjistit, který z těchto dvou detektorů lépe detekuje ticho vs. řeč. Sada testovacích skriptů byla dokumentována a společně s natrénovanými modely uložena do balíku výsledků testů.

Skutečné Indikátory dosažení - výsledky aktivity

Zpráva a balík výsledků ke stažení, s natrénovanými modely, okomentovanými testovacími skripty a dokumentací na adrese http://home.zcu.cz/~albrs/npv/ .

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Bylo zjištěno, že při přibližně stejných podmínkách vícevrstvý perceptron detekuje ticho vs. řeč u signálu s rušivými zvuky lépe než Gaussovské mixtury (92.0% vs. 88.5%). Testovací skripty, výpisy programů jsou k dispozici (ke stažení) – viz výše uvedená adresa.

Číslo aktivity

04

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Tvorba anotačních schémat

Zahájení aktivity

1.12.2006

Ukončení aktivity

30.6.2007

Popis aktivity

Anotační schéma přesně určuje všechny možnosti, jak lze větu sémanticky označit. Anotační schéma je hierarchická stromová struktura, každé anotační schéma definuje anotační značky pro téma, fráze, pod-fráze až k lexikálním třídám. Téma je kořenem stromu anotačního schématu a lexikální třídy jsou jeho listy. Fráze a pod-fráze pak tvoří uzly stromu, které nejsou ani listy, ani kořen. Anotační schéma bylo využito v aktivitě 2007-05 (sémantické anotování korpusu), jelikož podle anotačního schématu se věty anotovaly. V rámci řešení této aktivity byla nejprve vytvořena anotační schémata pro anotaci témat a vět a jejich vhodnosti pro další zpracování. Poté byla vytvořena další anotační schémata pro domény Předpovědi počasí, Městská hromadná doprava a Ubytování. Schémata pro další domény budou specifikována v pozdější fázi projektu. V rámci řešení této aktivity byly dále zdokonaleny jak editor anotací z aktivity 2006-02, tak metodologie vytváření sémantických anotací.

Skutečné Indikátory dosažení - výsledky aktivity

Byla vytvořena anotační schémata pro domény Předpovědi počasí, Městská hromadná doprava a Ubytování.
Editor sémantických anotací a navržená metodologie anotování vět byly prezentovány na významné konferenci Interspeech 2007 v publikaci: HABERNAL, I.; KONOPÍK, M. JAAE: The Java Abstract Annotation. In: Proceedings of Interspeech 2007, Bonn; ISCA, 2007, s. 1298-1301. ISSN 1990-9772.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Mezianotátorská shoda dosahovala 83% pro určení témat, 86% pro doménu Předpovědi počasí, 73% pro městskou hromadnou dopravu a 75% pro Ubytování. Shoda je dostatečná pro to, aby bylo možno prohlásit návrh anotačních schémat za úspěšný.

Číslo aktivity

05

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Sémantické anotování korpusu

Zahájení aktivity

8.1.2007

Ukončení aktivity

1.12.2007

Popis aktivity

V rámci této aktivity byl korpus získaný v aktivitě 2006-04 (Příprava a vytváření korpusu vět z vybraných domén) sémanticky označkován (určení významu vět). Sémantickou anotaci prováděl tým odborně vyškolených pracovníků sestavený v aktivitě 2006-07 (Výběr a zaškolení pracovníků provádějících sémantické anotace). Tým anotátorů pro svou práci využíval editor anotací získaný aktivitou 2006-02 (Vytvoření software pro editací sémantických anotací). Sémantické značkování probíhalo podle anotačních schémat navržených v aktivitě 2007-04 (Tvorba anotačních schémat). V první fázi této aktivity došlo k filtraci dotazů a k určení témat. V druhé fázi byl pro použitelnou část vět (počítačově zodpověditelných z informací na internetu) určen význam každé věty. Věta byla vždy anotována dvěma anotátory. Věty, ve kterých se anotátoři neshodli, byly dále analyzovány. Z analýzy vět rozdílně anotovaných byly provedeny změny v anotačních schématech (aktivita 2007-04) a v anotačním manuálu, podle kterého anotační pracovníci vytvářeli anotace. Pro všechny věty s rozdílnou anotací byl poloautomaticky určen jeden "správný" význam ve spolupráci s koordinátorem anotačních pracovníků (ing. Konopík). Touto aktivitou bylo vytvořeno dostatečné množství kvalitních trénovacích dat pro vývoj algoritmů automatické sémantické analýzy. Korpus bude nadále rozšiřován a následně anotován i po skončení této aktivity.

Skutečné Indikátory dosažení - výsledky aktivity

Bylo tématicky anotováno všech 20292 vět získaných v aktivitě 2006-04. Mezianotatátorská shoda dosahovala 83%. Následně bylo sémanticky anotováno 6750 vět s mezianotátorskou shodou 78%. Sémanticky byly anotovány věty z domén Předpovědi počasí, Městská hromadná doprava a Ubytování.
Anotační metodologie a vize celého projektu byly prezentovány na významné konferenci v Moskvě:
Konopík, M., Mouček, R.: Towards Semantic Analysis of Spoken Queries. In: SPECOM 2007 Proceedings, Moscow; Moscow State Linguistic University, 2007, pp. 817-822. ISBN 6-7452-0110-X.
Mouček, R., Konopík, M.: The Semantic Range of Spoken Dialogue Systems. In: SPECOM 2007 Proceedings, Moskva; Moscow State Linguistic University, 2007. pp. 720-724. ISBN 6-7452-0110-X.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Tématicky bylo anotováno 20292 vět. Sémanticky bylo anotováno 6750 vět. Toto množství je dostatečné pro vývoj algoritmů automatické sémantické analýzy. Anotace proběhly s dostatečnou mezianotátorskou shodou.
Sémantické anotace jsou uloženy na katedrálním serveru ve formátu XML a zálohovány na DVD.

Číslo aktivity

06

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj algoritmů pro automatickou identifikaci lexikálních tříd

Zahájení aktivity

1.9.2007

Ukončení aktivity

29.9.2008

Popis aktivity

Pojmem lexikální třída označujeme v této práci skupinu slov nebo slovních spojení, které dohromady spadají pod nějaký nadřazený pojem. Např. lexikální třída MĚSTA obsahuje mimo jiné názvy měst, např. Plzeň, Praha, Brno, ... Lexikální třída ČAS zahrnuje např. fráze "dvanáct hodin dvacet pět minut", "odpoledne", atd. Lexikální třídy jsou v korpusu anotovány v rámci jiné aktivity (2007-05: Sémantické anotování korpusu). V rámci této aktivity byly prozkoumány postupy automatické identifikace dané množiny lexikálních tříd v textu. Jako účinná metoda byla použita částečná sémantická analýza ručně vytvořenými bezkontextovými gramatikami pro lexikální třídy se složitější strukturou. Byl vytvořen nástroj založený na active bottom-up chart parseru, jehož koncept umožňuje následné sémantické vyhodnocení lexikálních tříd. Byly také prozkoumány metody založené na regulárních výrazech a slovníkovém přístupu.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem aktivity je sada algoritmů pro identifikaci lexikálních tříd v textu. Procentuální úspěšnost správné identifikace lexikální třídy je následující:
pro lexikální třídu "datum" 92%,
pro lexikální třídu "čas" 94% a
pro lexikální třídu "čísla" 98%.
Trénování a měření byla prováděna na testovacích větách z aktivity 2007-05.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky práce byly prezentovány a obhájeny v diplomové práci: Habernal I.: Lexical Class Analysis, ZČU, 2007; text práce je k nalezení v přílohách.

Číslo aktivity

07

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Ověření vlastností vhodného typu neuronové sítě pro zpracování sémantiky přirozeného jazyka

Zahájení aktivity

2.1.2007

Ukončení aktivity

22.12.2007

Popis aktivity

Cílem této aktivity bylo zvolit vhodný typ umělé neuronové sítě pro zpracování sémantiky přirozeného jazyka a ověřit její vlastnosti. Z podrobnější analýzy článků a dostupných informací a na základě vlastností jednotlivých typů umělých neuronových sítí, se kterými jsme v minulosti pracovali a měli určité zkušenosti, byla nakonec pro zpracování sémantiky zvolena Kohonenova samoorganizující mapa. Důvody pro její volbu byly následující: • Jedná se o relativně jednoduchou architekturu, která má poměrně široké spektrum využití. • Existují kvalitní volně dostupné simulátory, takže pro ověření vlastností není nutné v první fázi provádět vlastní implementaci. • Tato síť byla k řešení podobného problému využita v systému WEBSOM, který zpracovával sémantiku anglického a finského jazyka a umožňoval efektivní prohledávání dokumentů a vyhledávání odpovídajících informací. Pro počáteční testování vlastností sítě byly k dispozici volně dostupné simulátory SOMPAK (navržený pro prostředí MSDOS) a SOMtoolbox (navržen pro prostředí MatLab). Testy ukázaly, že oba simulátory jsou využitelné pro

řešení daného problému, SOMPAK však nelze použít pro zpracování velkého množství dokumentů, neboť neumožňuje (z důvodů implementace pro prostředí MSDOS) vytvářet rozsáhlé mapy. Dalším úkolem, který bylo třeba vyřešit, byla volba vhodné metriky, která by byla využitelná v procesu shlukování jednotlivých slov, popř. celých kolekcí česky psaných dokumentů. Kohonenova mapa byla původně navržena pro zpracování číselných vstupních vektorů. Z tohoto důvodu se pro určování míry podobnosti vstupních vektorů využívá euklidovská metrika. Pokud bychom chtěli tuto metriku využít pro výše uvedenou úlohu, bylo by nutné vhodným způsobem transformovat textovou informaci do číselné podoby. Další možností je modifikovat Kohonenovu mapu tak, aby umožňovala přímo zpracovávat textovou informaci. V tomto případě by bylo možné využít např. Levensteinovu metriku, často využívanou pro určování míry podobnosti mezi textovými řetězci. Druhý uvedený postup byl důkladně analyzován a byl porovnáván s již zmíněnou transformací textové informace do číselné podoby. Ukázalo se, že přímé použití Levensteinovy metriky pro posuzování podobnosti slov je nevhodné z následujících důvodů: • Slova, která jsou si vzájemně morfologicky podobná (a jejich Levensteinova vzdálenost je malá), mohou mít zcela odlišný význam, což by v konečné fázi značně ovlivnilo posuzování podobnosti u celých dokumentů. • Kohonenovu mapu, kterou budeme pro posuzování podobnosti využívat, bude nutné upravit tak, aby mohla uchovávat textové vektory, což je opět problém, poněvadž by jako váhové vektory sítě bylo nutné uchovávat celé textové řetězce, popř. seznamy textových řetězců. • Inicializace Kohonenovy mapy textovými řetězci by byla komplikovaná a výsledek trénování mapy textovými řetězci je v tomto případě nepredikovatelný.

Skutečné Indikátory dosažení - výsledky aktivity

Výběr neuronové sítě - Kohonenova samoorganizující se mapa a následné vytvoření mapy slovních kategorií.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byl zpracován programový balík stažitelný ze serveru http://liks.fav.zcu.cz/mediawiki/index.php/WEBSOM.
Uveřejněno pod licencí GPL v.2, dokumentace k programovému balíku byla zpracována do podoby textového souboru (Mouček, R., Mautner P.: Projekt WEBSOM) uloženého tamtéž.
Výsledky zveřejněny v článku Mautner P., Mouček R.: Zpracování česky psaných dokumentů Kohonenovou samoorganizující mapou, Sborník 3. ročníku konference s mezinárodní účastí Informatika v škole a praxi, Ružomberok, Slovensko, 2007

Číslo aktivity

08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření vhodných vstupů neuronové sítě

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.9.2007

Popis aktivity

Přímé zpracování textové informace Kohonenovou mapou je nevhodné (vstup Kohonenovy mapy a výpočty v mapě jsou založeny na numerickém základě) a je tedy nutné nalézt vhodnou transformaci textových řetězců do číselné podoby (použity byly zatím textové řetězce z článků databáze ČTK). Pro transformaci byl použit postup s algoritmem WEBSOM (parametrizovaný). Tento postup umožňuje kódovat jednotlivá slova ze slovníku jako číselné vektory. Jednotlivé číselné komponenty v těchto vektorech zároveň reprezentují kontext, v jakém se slova vyskytují v textovém dokumentu. Na základě takto kódovaných slov je trénováním Kohonenovy mapy vytvořena tzv. mapa slovních kategorií, tj. vytvoří se shluky slov, které se v dokumentech vyskytují v podobném kontextu. Podrobně je tento způsob uveden v citovaném článku. V současnosti probíhají další testy (modifikace podoby vstupních vektorů) tak, aby tyto lépe odrážely sémantický kontext dokumentů.

Skutečné Indikátory dosažení - výsledky aktivity

Číselná reprezentace textů - dokumentů z databáze ČTK - tvoří vstupy Kohonenovy samoorganizující se mapy + slovník (lemmatizovaný) z dokumentů databáze ČTK

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byl zpracován programový balík stažitelný ze serveru http://liks.fav.zcu.cz/mediawiki/index.php/WEBSOM.
Uveřejněno pod licencí GPL v.2, dokumentace k programovému balíku byla zpracována do podoby textového souboru (Mouček, R., Mautner P.: Projekt WEBSOM) uloženého tamtéž.
Výsledky zveřejněny v článku Mautner P., Mouček R.: Zpracování česky psaných dokumentů Kohonenovou samoorganizující mapou, Sborník 3. ročníku konference s mezinárodní účastí Informatika v škole a praxi, Ružomberok, Slovensko, 2007

Číslo aktivity

09

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Návrh a implementace modelu webgrafu s cílem určit autoritativnost uzlů zohledňující skryté vazby komponent

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.6.2007

Popis aktivity

Data z uznávané digitální knihovny DBLP ve formátu XML byla převedena do relační databáze a na ni byly aplikovány algoritmy s modifikovaným vzorcem pro výpočet autorit. Tyto metody jsme využívali již v předchozí fázi projektu, ovšem data pocházela přímo z webu a nebyla proto zcela přesná a spolehlivá. To nám tedy neumožňovalo srovnávat výsledky s jinými relevantními údaji. Naproti tomu využití zdrojů DBLP takové srovnání velmi usnadňuje a dosažené výsledky jsou velmi povzbudivé. Digitální knihovna DBLP (http://dblp.uni-trie.de) v současné době obsahuje přes 900 tisíc bibliografických záznamů o nejrůznějších publikacích v oblasti databází, digitálních knihoven, softwaru, hardwaru, data miningu, algoritmů, paralelních procesů apod. Pro použití dat z této knihovny jsme ji museli z původního formátu (jednoho souboru XML) převést do relační databáze SQLite. Tu jsme zvolili kvůli její jednoduchosti a volné dostupnosti na webových stránkách (http://www.sqlite.org). Software v jazyce C# vyvinutý speciálně pro tento převod je ke stažení na adrese http://home.zcu.cz/~dalfia/CotSewing/DBLP.zip. Pro oblast bibliografických sítí a zjišťování autoritativnosti autorů odborných publikací jsme navrhli sedm alternativních variant známého algoritmu PageRank pro určování důležitosti webových stránek. Tyto varianty se od klasického PageRanku odlišují především tím, že berou v potaz nejen citace mezi autory, ale i jiné informace obsažené v bibliografických sítích, např. informace o počtu společných publikací dvou citujících se autorů, o počtu spoluautorů ve společných publikacích atd. Hlavní motivací těchto modifikací PageRanku je snaha učinit hodnocení autorů „férovějším“ a využít doplňující informace, které se na webu nenacházejí, ale v bibliografických sítích ano. Ústřední myšlenkou je zde předpoklad, že všechny citace mezi autory nemají stejnou váhu. Např. citace dvou často spolupracujících autorů – kolegů – je zřejmě méně hodnotná než citace mezi dvěma autory, kteří se neznají. Zdrojové texty programů implementujících výše uvedené metody jsou k dispozici na http://home.zcu.cz/~dalfia/CotSewing v souborech Aggregator.zip (příprava databáze pro výpočet) a RankCalculator.zip (samotné implementace algoritmů pro výpočet důležitosti).

Skutečné Indikátory dosažení - výsledky aktivity

Zřejmým indikátorem dosažení je funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy. Důležitá je rovněž možnost různých relevantních nastavení. Zdrojový kód výsledné aplikace musejí být řádně dokumentovány a jeho modulární členění musí umožnit jeho snadnou údržbu a pozdější rozšiřování.
Jednoznačným indikátorem dosažení je existence relační databáze s daty z digitální knihovny DBLP původně uloženými ve formátu XML. Logická struktura takovéto databáze musí být vhodná pro efektivní získávání informací z DBLP, tj. musí být navržena pro jednoduchou tvorbu optimálních dotazů v SQL.
Nejdůležitějším výsledkem je návrh a implementace algoritmů pro hodnocení důležitosti uzlů v orientovaném grafu. Tímto grafem může být např. graf webových stránek nebo citační graf autorů odborných publikací apod. Výše

uvedené algoritmy však byly „ušity na míru“ bibliografickým sítím. Výsledky aplikací celkem 12 algoritmů (vedle zmiňovaných sedmi variant PageRanku také klasický PageRank, vážený PageRank, HITS, vážené a nevážené citace) na data knihovny DBLP ve verzi z února 2004 jsou dostupné v tabulkách databáze dblp.db (ve formátu SQLite), jež je zájemcům k dispozici u autora (dalfia@kiv.zcu.cz).

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Veškeré výsledky a závěry byly prezentovány v pracích a článcích:

Fiala D.: Web Mining Methods for the Detection of Authoritative Sources. PhD thesis, University of West Bohemia in Pilsen, Czech Republic and University Louis Pasteur – Strasbourg I, France, 2007.

Fiala D., Rousselot F., Ježek K.: PageRank for Bibliographic Networks, Scientometrics, Vol. 76, No. 1, 2008 (přijato k publikaci).

Kopie textu článku a disertační práce lze nalézt v příloze této zprávy.

Číslo aktivity

10

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Vyhodnocení výsledků navržených algoritmů pro analýzu struktury Webu

Zahájení aktivity

2.5.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Bylo zapotřebí provést rozsáhlé srovnávání výsledků našich algoritmů s jinými, již existujícími informacemi. Nejprve bylo nutno vymyslet a navrhnout srovnávací testy. Náš webový pavouk „prošel“ sedmnáct serverů českých kateder informatiky a osmdesát serverů francouzských kateder informatiky. Informace o odkazech mezi stránkami uložil do databáze a vytvořil korpus stažených dokumentů potřebný pro další analýzu. Přitom nás zajímaly pouze odkazy přes protokol HTTP k dokumentům v určitých formátech. Například nebyly brány v úvahu audio a video soubory, což je přirozené, ale také byly vynechávány dokumenty s příponami doc, rtf, txt a ppt, což už může být označeno jako spornější. Aby bylo zabráněno pavoukovi uvíznout v nějaké webové pasti, byla stanovena hloubka osm jako maximální hloubka vnoření do webového grafu, což je podle dosavadních zkušeností dobrý odhad pro získání rozumných výsledků. Zdrojový kód stahovacího programu v jazyce C# je k nalezení na adrese http://home.zcu.cz/~dalfia/WebWatch3.zip. Kromě zkoumání odkazů ve skupině webových domén kateder informatiky jsme se rovněž zabývali analýzou samotných dokumentů na těchto serverech. Mimo soubory obsahující odkazy (především HTML stránky) byly také stahovány potenciální odborné články. V praxi to znamenalo shromažďovat dokumenty PDF a PS (PostScript), protože většina odborných publikací veřejně dostupných na webu je v těchto dvou formátech. Dalším úkolem bylo extrahovat z článků informace potřebné pro citační analýzu – jména autorů, názvy článků, atd. Zde byla aplikována stejná metoda jako v [10], [12] založená na skrytých Markovových modelech (SMM). Z citačního grafu, kde uzly byly „příjmení“, byly třemi různými hodnoticími metodami (In-Degree, HITS, PageRank) určeni nejautoritativnější čeští a francouzští autoři (rozpoznání českých a francouzských příjmení bylo prováděno ručně). Zdrojové texty programů implementujících výše uvedené metody jsou k dispozici na http://home.zcu.cz/~dalfia/CotSewing v souborech Aggregator.zip (příprava databáze pro výpočet) a RankCalculator.zip (samotné implementace algoritmů pro výpočet důležitosti).

Skutečné Indikátory dosažení - výsledky aktivity

Důležitým výsledkem bylo vytvoření webového grafu českých a francouzských kateder informatiky. Jejich grafická podoba je k dispozici na adrese http://home.zcu.cz/~dalfia/papers/ v souborech Czech.svg a France.svg. Na základě těchto dvou grafů bylo několika různými metodami stanoveno pořadí českých a francouzských vědců v oboru informatiky podle autoritativnosti. Jako srovnávací postup bylo možno použít porovnání žebříčků se seznamy vědců oceněných organizací ACM, např. s nositeli ocenění ACM SIGMOD E. F. Codd Innovations Award (http://www.sigmod.org/sigmodinfo/awards), čehož jsme využili v předcházející aktivitě č.9.

Podrobné informace lze najít ve výše uvedených publikacích a také v disertační práci citované v popisu aktivity 9.
Dalším verifikovatelným výsledkem aktivity je návrh metody testování kvality algoritmů pro hodnocení webových stránek. a její implementace. Algoritmy navržené v aktivitě č. 10 pro rok 2007 a v aktivitách č. 15 a č. 16 pro rok 2006 bylo nutné otestovat na snadno dostupných datech a stanovit jasná kritéria pro porovnání úspěšnosti jednotlivých metod. Testovací metodologie musela být jednoznačná a reprodukovatelná. Na jejím základě muselo být možno stanovit, které algoritmy generují lepší žebříčky důležitosti uzlů v orientovaném grafu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky a závěry byly prezentovány v článcích:
Fiala D., Rousselot F., Ježek K. Ranking Algorithms For Web Sites: Finding Authoritative Academic Web Sites and Researchers. Proceedings of the 3rd International Conference on Web Information Systems and Technologies WEBIST'07, Barcelona, Spain, pp. 372-375, 2007.
Fiala D., Ježek K., Rousselot F. Využití struktury webu pro vyhledávání autoritativních institucí a osob. Proceedings of the 6th Annual Conference ZNALOSTI 2007, Ostrava, Czech Republic, pp. 300-303, 2007.
Kopie textů článků lze nalézt v příloze této zprávy.

Číslo aktivity

11

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

SPOT – nový webový projekt on-line slovníku překladů odborných termínů

Zahájení aktivity

3.1.2007

Ukončení aktivity

20.12.2007

Popis aktivity

Na základě analýz a podnětů zájemců o projekt z roku 2006 pokračovaly práce na implementaci v rámci jedné diplomové práce (P. Klesová, vedoucí Jiří Hynek) a oborového projektu (O. Čabrada, vedoucí Přemysl Brada). Výsledkem byl upgrade aplikace s doplněnou následující podstatnou funkčností, která přispívá k tvorbě komunity kolem slovníku: • hlasování o kvalitě překladu, • zobrazení kontextu slova (Google search, oborové weby, blogy), • podpora pro týmy (překladatelské projekty), • internacionalizace aplikace a lokalizace do CZ, EN, DE. Dále byl zprovozněn informační server http://blogspot.zcu.cz/ poskytující prostor pro publikování poznatků, názorů a aktuálních informací o projektu. Další rozvoj slovníku je zajištěn probíhajícím vývojem, zatím v rámci dvou bakalářských prací. Cíle pro nejbližší aktualizaci aplikace jsou zvýšit uživatelský komfort a doplnit dávkovou aktualizaci slovní zásoby uploadem souboru s korpusem. Detaily jsou dostupné z wiki stránek projektu http://wiki.kiv.zcu.cz/SlovnikTerminologie/HomePage.

Skutečné Indikátory dosažení - výsledky aktivity

• Funkční aplikace dostupná na http://spot.zcu.cz/ naplněná iniciačním korpusem. Ověřovací provoz s pilotní množinou uživatelů.
• Doprovodný blog k výše uvedenému na adrese http://www.blogspot.cz (cca 6000 přístupů v prvním měsíci po uvedení do provozu).
• Obhájena diplomová práce Petry Klesové: Slovník oborové terminologie (2007)

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku. Výsledky byly publikovány v
• Hynek, J., Brada, P.: On the Evolution of Computer Terminology and the SPOT on-Line Dictionary Project. In: Openness in Digital Publishing: Awareness, Discovery and Accesss; Vienna, ÖKK-Editions, 2007, s.257-268. ISBN 978-3-85437-292-9
• Hynek, J.: Praktický technický slovník anglicko-český / česko-anglický; Fraus, Plzeň, Czech Republic, 2007, ISBN 978-80-7238-640-6

Číslo aktivity

12

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava sumarizačních kolekcí – multi-dokument sumarizace v češtině

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.11.2007

Popis aktivity

Z důvodu neexistence kolekce českých dokumentů anotovaných pro sumarizaci byl vytvořen sumarizační korpus. Tento korpus obsahuje shluky dokumentů. Každý shluk se týká určité události/tématu. Navíc každý shluk obsahuje 3 dotazy/úlohy. Anotátoři vytvořili jak abstrakty (souhrnem je text napsaný anotátorem), tak extrakty (souhrnem jsou vybrané věty původních dokumentů) pro každý shluk a dotaz. Vše je ukládáno v XML formátu.

Skutečné Indikátory dosažení - výsledky aktivity

Korpus je v XML formátu volně stažitelný na stránkách textmining.zcu.cz . V současné době obsahuje 7 shluků dokumentů, 71 dokumentů. V každém shluku jsou 3 dotazy – jeden obecný, jeden specifický a jeden inkrementální. Pro každý shluk a dotaz jsou k dispozici abstrakty a extrakty vytvořené čtyřmi anotátory. V další etapě plánujeme rozšíření korpusu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Korpus byl použit pro testování multi-document sumarizátoru založeném na LSA (viz aktivita 13). Výsledky byly publikovány v:

1. Steinberger, J., Křišťan, M: LSA-Based Multi-Document Summarization. In: Proceedings of 8th International PhD Workshop on Systems and Control, Balatonfured, Maďarsko, 2007.

2. Steinberger, J, Tesař, R.: Knowledge-poor Multilingual Sentence Compression. In: Proceedings of the seventh conference on language engineering, Egyptian Society of Language Engineering, Káhira, Egypt, 2007.

Číslo aktivity

13

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Experimentální verze sumarizátoru založeného na LSA – multi-dokument sumarizace

Zahájení aktivity

2.1.2007

Ukončení aktivity

21.12.2007

Popis aktivity

V předešlé etapě projektu byla latentní sémantická analýza úspěšně aplikována na případ single-document sumarizace (vstupem je pouze jeden dokument). Cílem této etapy řešení sumarizačního problému bylo aplikování podobných postupů na problém multi-document sumarizace (vstupem je shluk dokumentů, vztahujících se ke stejné události/tématu). Souhrn lze navíc řídit uživatelským dotazem. Také byla zkoumána možnost komprese vět v extraktu a využití rezoluce anafor. Okrajově se řešitel také zabýval možnostmi vyhledávání souvisejících informací ve wikipedii a rozhodování, zda tyto věty obsahují pro dané téma (téma bylo dáno jedním konkrétním článkem) nové nebo redundantní informace. Tato část bude využita při tvorbě inkrementálních souhrnů.

Skutečné Indikátory dosažení - výsledky aktivity

Byl vytvořen systém, který vytváří souhrny shluků dokumentů. Popsán byl v publikaci [1]. Dále byl vytvořen modul pro využití rezoluce anafor a pro kompresi souvětí. Problematika související s využitím rezoluce anafor při

sumarizaci byla popsána v publikaci [2]. Metoda a pilotní experimenty s kompresí vět v souhrnu byly popsány v publikaci [3]. Z projektu byla také částečně financována práce na přípravě finální verze článku zabývajícího se experimenty s wikipedií (vyhledávání souvisejících nových informací) [4]. Dále řešitel obhájil disertační práci na téma Sumarizace textů založená na LSA [5].

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky aktivity byly prezentovány v následujících publikacích:

[1] Steinberger, J., Křišťan, M.: LSA-Based Multi-Document Summarization. In: Proceedings of 8th International PhD Workshop on Systems and Control, Balatonfured, Maďarsko, 2007.

[2] Steinberger, J., Poesio, M., Kabadjov, M.A., Ježek, K.: Two uses of anaphora resolution in summarization. In: Information Processing and Management, 43(6), Elsevier Ltd., 2007.

[3] Steinberger, J., Tesař, R.: Knowledge-poor Multilingual Sentence Compression. Proceedings of the seventh conference on language engineering, Egyptian Society of Language Engineering, Káhira, Egypt, 2007.

[4] Sutcliffe, R.F.E., Steinberger, J., Kruschwitz, U., Kabadjov, M.A., Poesio, M.: Identifying Novel Information Using Latent Semantic Analysis in the WiQA Task at CLEF 2006. In: Lecture Notes in Computer Science, No. 4730, Springer Verlag, Berlin, Heidelberg, 2007.

[5] Steinberger, J.: Text Summarization within the LSA Framework. Disertační práce, ZČU Plzeň, 2007.

Číslo aktivity

14

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Metoda hodnocení kvality sumarizátorů na základě LSA

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.9.2007

Popis aktivity

V oblasti hodnocení kvality sumarizace textů se stále hledají nové automatické metody. Jediná všeobecně uznávaná automatická metoda ROUGE dosahuje slušných výsledků, avšak stále nedostačujících, a navíc potřebuje mít vytvořeny abstrakty k jednotlivým dokumentům/shlukům dokumentů. Cílem této aktivity bylo nalezení způsobu využití LSA pro hodnocení kvality sumarizátorů. Myšlenkou navržené metody je, že hlavní témata (vektory získané z LSA) souhrnu by měly být co nejvíce podobná hlavním tématům originálního textu. Referenčním dokumentem může být i abstrakt. Experimenty ukázaly vhodné schéma vážení termů. Navržená metoda koreluje lépe s hodnocením anotátorů v porovnání s baseline metodami v případě použití plného textu jako referenčního dokumentu. V případě použití abstraktů je lepší ROUGE.

Skutečné Indikátory dosažení - výsledky aktivity

V rámci řešení byl vytvořen experimentální systém, který umožňuje spuštění různých hodnotících metod, včetně LSA metody. Metoda a výsledky byly publikovány v [1].

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Publikace:

[1] Steinberger, J., Ježek, K.: Evaluation Measures in Text Summarization. Přijato do časopisu Computing and Informatics, Slovenská akademie věd, 2007/2008.

Číslo aktivity

15

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Aplikace pro anotaci textů a hodnocení sumarizačních metod

Zahájení aktivity

2.1.2007

Ukončení aktivity

21.12.2007

Popis aktivity

V předchozí etapě projektu byla vytvořena aplikace pro anotaci single-document sumarizace. S přechodem k tvorbě multi-document souhrnů bylo nutné tuto aplikaci rozšířit o možnost tvorby souhrnů shluků dokumentů – abstrakty i extrakty. Dílčí úlohy řešené při sumarizaci – komprese souvětí a kontrola referencí souhrnů potřebují pro možnost porovnání s anotátory také patřičné anotace, o které byl systém dále rozšířen. Druhá část systému obsahuje možnosti spuštění různých sumarizačních metod, hodnocení kvality jejich výstupů různými hodnotícími metodami a výpočet shody anotátorů.

Skutečné Indikátory dosažení - výsledky aktivity

V rámci této aktivity byl rozšířen systém pro anotaci sumarizačního korpusu o možnost tvorby abstraktů/extraktů shluků dokumentů (multi-document sumarizace), o anotaci komprese vět v extraktech a anotaci nahrazování referencí v extraktech. Systém byl použit pro tvorbu českého sumarizačního korpusu (viz aktivita 12).

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Aplikace byla zpracována v jazyce Java a je volně stažitelná ze stránek textmining.zcu.cz.

Číslo aktivity

16

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření systému pro vyhledávání

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

V rámci projektu navrhujeme prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jádrem vyhledávání je thesaurus EuroWordNet a sumarizátor je založen na latentní sémantické analýze. V současném řešení se zaměřili na zpracování anglického a českého jazyka. Jelikož princip zpracování zůstává stejný i pro ostatní jazyky poskytované tezaurem EWN, plánujeme vyzkoušet postup i pro další evropské jazyky. Hlavní důraz bude kladen na zvýšení relevance vyhledaných dokumentů. Aplikace se zdrojovými kódy je k dispozici na www.kiv.zcu.cz. Součástí jsou také korpusy obsahující testovací data. Program je připraven k použití na datech předzpracovaných pomocí tezauru EWN. Převodní programy a potřebné lemmatizátory jsou součástí řešení. Pro použití je nutné postupovat v jednotlivých krocích: 1) příprava lemmatizačních slovníků pomocí Ispellu, 2) mapování lemmat na indexy EWN, 3) Import korpusů do systému pro vyhledávání a sumarizaci, 4) Import lemmat a EWN slovníků do systému, 5) Vyhledávání a verifikace. Verifikace je prováděna ručně v porovnání s výsledky metod Gogole Desktop Search. Pro tento test se rozdělí testovací korpusy na články uložené do vyhrazeného adresáře, který je zpracován algoritmem ve vyhledávači Google. Srovnání se provede standardním způsobem pro výpočet přesnosti a úplnosti.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy. Aplikace je vytvořena v programovacím jazyku Java, skládá se z modulu pro vyhledávání založeném na vektorovém modelu, sumarizaci pomocí LSA a modulu bayesovského disambiguátoru. Programy jsou k dispozici ke stažení z webové stránky katedry - www.kiv.zcu.cz spolu s dokumentací a daty. Výsledky aktivity budou použity v rámci řešení aktivity pro rok 2008 "Experimenty v systému pro vyhledávání".

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku na vybraných datech z aktivity 19. Vyhodnocována byla přesnost a úplnost výsledků vyhledávání.

Číslo aktivity

17

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Rozšiřování uživatelského dotazu

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

V rámci projektu jsme navrhli prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jednou s možností vylepšující relevanci dokumentů vyhledávaných uživatelem je rozšíření dotazu. Touto progresivní oblastí se hodláme zabývat a očekáváme zlepšení relevance a pokrytí výsledků hledání. Aplikace se zdrojovými kódy je k dispozici na www.kiv.zcu.cz. Součástí jsou také korpusy obsahující testovací data. Program je připraven k použití na datech předzpracovaných pomocí tezauru EWN. Prováděné testy lze rozdělit do dvou skupin: 1) Monolinguální testy – dotazy jsou zadány v libovolném jazyku a následně jsou vyhledány články relevantní dotazu v daném jazyce. 2) Testy napříč jazyky – hledaný term je zadán v jazyku A a vyhledávány jsou dokumenty v jazyku B, přičemž jazyky A a B jsou odlišné. Rozšíření dotazu je v danou chvíli možné následujícím způsobem: 1) Synonyma 2) Hypernyma 3) Holonyma Verifikace je prováděna ručně pro každý vyhledaný článek korpusu. Srovnání se provádí standardním způsobem pro výpočet přesnosti a úplnosti.

Skutečné Indikátory dosažení - výsledky aktivity

Navržený systém obsahuje možnost rozšiřování uživatelských dotazů s použitím tezauru EWN. Výsledky byly vyhodnoceny standardními statistickými metodami, sledována byla především přesnost a úplnost vyhledávání. Další testy budou provedeny v následujícím roce v rámci plánované aktivity pro rok 2008 "Experimenty v systému pro vyhledávání".

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Systém byl testován na relevanci získávaných výsledků jak pro české a anglické prostředí, tak i při křížovém zpracování. Porovnání bylo provedeno ručně zaškolenými evaluátory. Programy a použitá data jsou k dispozici na stránkách www.kiv.zcu.cz spolu s dokumentací.

Číslo aktivity

18

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Metody disambiguace ve vyhledávací úloze

Zahájení aktivity

2.1.2007

Ukončení aktivity

1.9.2007

Popis aktivity

Pro aplikaci jsme použili jednu z metod využitelných pro disambiguaci – bayesovský disambiguátor. Tato metoda rozlišuje významy slov na základě kontextu, ve kterém se vyskytují. Ve standardní verzi se kontext považuje za neuspořádanou množinu slov. Pro disambiguaci jsme provedli převedení korpusu do indexů EWN. Pro trénování disambiguárotu jsme použili Brownův korpus, který obsahuje označkované slova s odpovídajícími čísly synsetů,

což jsou synonymické množiny tezauru EWN. Postup zpracování probíhá následujícím způsobem: 1) Zadání dotazu uživatelem 2) Lemmatizace dotazu 3) Disambiguace bayesovskou metodou 4) Převod na indexy EWN 5) Vyhledání 6) Alternativně může dojít k rozšíření dotazu podle aktivity 17. Jelikož je navržený systém multilinguální, je možné provádět disambiguaci libovolného jazyka obsaženého v tezauru EWN.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikační modul schopný provádět disambiguaci slov v anglickém a českém jazyce v souvislosti s vyhledávací úlohou. Výsledky byly vyhodnoceny standardními statistickými metodami, sledována byla především přesnost a úplnost vyhledávání v porovnání s metodou, kde není disambiguace použita. Další testy budou provedeny v následujícím roce v rámci plánované aktivity pro rok 2008 "Experimenty v systému pro vyhledávání".

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky systému jsou testovány na vybraných dokumentech a porovnány s očekávaným výsledkem. Evaluace byla provedena zaškolenými pracovníky. Programy a použitá data jsou k dispozici na stránkách www.kiv.zcu.cz spolu s dokumentací.

Číslo aktivity

19

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Klasifikační metody v multijazykovém prostředí

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Vzhledem k rostoucímu významu mezinárodní kooperace nabylo na důležitosti zpracování dokumentů v různých jazycích. Proto bylo rozhodnuto rozšířit klasifikační metody a modifikovat pro použití v multilingválním prostředí. V experimentech byla sledována míra ovlivnění výsledků klasifikace normalizací slov a odstraněním stop-slov na dvou korpusech. Pro testy byl vybrán klasifikátor multinomial Naive Bayes. Kvalita klasifikace byla porovnávána standardními metrikami, jmenovitě mírou micro-F1, macro-F1, přesností a úplností. Statistická významnost výsledků byla vyhodnocena použitím t-testu a dále byla použita technika 4-cross fold validace. Oba korpusy byly nejdříve předzpracovány rozdílnými normalizačními technikami a následně klasifikovány. Pro experimenty jsme použili dva textové korpusy - anglické dokumenty vybrané z Reuters Corpus Volume 1 a české dokumenty agentury ČTK. Testovali jsme 6 normalizačních algoritmů – Lovins, Iterated Lovins, Paice, Porter's stemmer, EWN metodu s použitím a bez použití synsetů. Pro český jazyk byla do testů zahrnuta také algoritmická lemmatizace. Tato aktivita souvisi s aktivitou 22, kde jsou použity shodné algoritmy a vstupní data.

Skutečné Indikátory dosažení - výsledky aktivity

Byla provedena klasifikace dokumentů lemmatizovaných různými metodami a porovnán jejich vliv na kvalitu výsledných korpusů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Ověřili jsme vliv lemmatizace na klasifikační úloze. Použili jsme klasifikátor Multinomial Naive Bayes a srovnali výsledky s ostatními přístupy. V souladu s plánem byla vytvořena vlastní testovací kolekce z dat ČTK (České tiskové kanceláře) v českém jazyce, na které bylo provedeno otestování vlivu různých přístupů k normalizaci slov na úspěšnost klasifikace. Testovací data a potřebné aplikace jsou k dispozici webových stránkách www.kiv.zcu.cz. Aplikace je připravena k použití.

Výsledkem akce je publikace:

• Toman, M., Tesař, P., Ježek, K.: Vliv normalizace slov na klasifikaci textů. Publikováno na konferenci Znalosti 2007, Ostrava: VŠB - Technická univerzita, ISBN 978-80-248-1279-3, s. 360-363, únor 2007.

Číslo aktivity

20

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Doplnění jazykových korpusů o další evropské jazyky

Zahájení aktivity

2.1.2007

Ukončení aktivity

1.7.2007

Popis aktivity

Vzhledem k rozšiřující se mezinárodní kooperaci nabývá na důležitosti zpracování dokumentů v různých jazycích. Proto jsme se rozhodli datové korpusy rozšířit o další evropské jazyky.

Skutečné Indikátory dosažení - výsledky aktivity

Při řešení projektu byla vytvořena multilinguální databáze obsahující dokumenty ve významnějších evropských jazycích – většina korpusu je v českém a anglickém jazyce. Jeden korpus je multilinguální obsahuje cca 2000 článků v hlavních evropských jazycích. Na těchto korpusech bylo následně prováděno testování vytvářených algoritmů a navrhovaných postupů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

U vytvořeného korpusu byla provedena vizuální kontrola a filtrace vhodných článků pro zamýšlené testy. Z 200 vybraných článků byly vytvořeny ruční extrakty, které poslouží jako podklad pro následnou aktivitu roku 2008 "Experimenty v systému pro vyhledávání". Byly vytvořeny korpusy z webových zdrojů washingtonpost.com, dw-world.de, reuters.com, nytimes.com, news.bbc.co.uk, iht.com, telegraph.co.uk, timesonline.co.uk, independent.co.uk, forbes.com, usatoday.com, guardian.co.uk, pravo.cz, blesk.cz, lidovky.cz, mfdnes.cz, novinky.cz, idnes.cz, ct24.cz; tedy 7 českých, 14 anglických, 1 multilinguální. Korpusy s jejich popisem jsou k dispozici na webových stránkách www.kiv.zcu.cz.

Číslo aktivity

21

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření rozhraní ke klasifikátoru SVM

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.6.2007

Popis aktivity

Klasifikátor SVM je v současné době jediný klasifikátor, který dosahuje vynikajících výsledků prakticky ve všech oblastech zpracování dat, včetně zpracování textu. Stalo se proto jednou z priorit tohoto projektu využít jeho vlastností a prozkoumat jeho efektivitu – například při použití různých modelů charakterizujících textové dokumenty. Protože je již na internetu k dispozici jeho univerzální implementace (viz http://svmlight.joachims.org), bylo v rámci této aktivity vytvořeno základní aplikační rozhraní, které umožňuje použití různých technik využívaných při zpracování textu současně s klasifikátorem SVM. Účelem rozhraní je zjednodušit použití klasifikátoru SVM pro potřeby zpracování textu a odstranit některé kroky, které musejí být díky snaze o co jeho nejuniverzálnější použití v různých oblastech prováděny. Byl vytvořen funkční základ potřebný pro implementaci dalších nadstaveb a do rozhraní byla zakomponována možnost použít k jednotlivým slovům reprezentujícím textový dokument také bigramy a 2-itemsety. V minulém období byl také objeven úspěšnější přístup k převodu reprezentace textového dokumentu do podoby akceptované SVM klasifikátorem. Popis aplikace a jejího použití je k dispozici na katedře v zip archivu současně se zdrojovými kódy. Rozhraní je připraveno k použití. Bylo vytvořeno proto, aby akceptovalo

vstup textových dokumentů ve franta-formátu a umožňovalo reprezentovat textové dokumenty kromě samostatných slov i bigramy a itemsety, a to i současně. Zpracování navazuje na aktivitu číslo 18/2006.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat definované vstupy a poskytnout požadované výstupy.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity

22

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Implementace klasifikační metody Naive Bayes rozšířené o současné zpracování bigramů a 2-itemsetů

Zahájení aktivity

1.7.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Vlastnosti metody Naive Bayes a její úspěšnost při klasifikaci textových dokumentů byly poměrně podrobně prozkoumány. Stále se zde však objevovaly nové prostory pro zkoušení nových přístupů vedoucích ke zlepšení dosahovaných výsledků. Jedním z nich byla i možnost obohatit obecně používaný bag-of-words model dokumentu (tedy model založený na jen jednotlivých slovech) o další položky. Těmi mohou být například slovní n-gramy nebo itemsety. Doposud byly n-gramy a itemsety zkoumány vždy odděleně, naším cílem se však stalo nejen ověřit, které z obou přístupů poskytují lepší výsledky, ale navíc se i pokusit oba přístupy k zlepšení úspěšnosti klasifikace zkombinovat za účelem vylepšení dosavadních výsledků. A právě uskutečněním této aktivity se otevřela možnost tyto experimenty realizovat. Díky dokončení této aktivity jsme již měli možnost dosáhnout poměrně zajímavých výsledků. V uplynulé době byly dokončeny veškeré kroky nutné k použití bigramů a 2-itemsetů současně k obohacení modelů textových dokumentů. V souladu s naším plánem byla vytvořena i vlastní testovací kolekce z dat ČTK v českém jazyce, která doposud chyběla a na které jsme provedli otestování vlivu různých přístupů k normalizaci slov na úspěšnost klasifikace. Nad rámec původního plánu byl vytvořen univerzální nástroj Teraman, který slouží k extrakci n-gramů z rozsáhlých textových kolekcí. Tento nástroj je schopen extrahovat libovolný řád n-gramů z kolekcí čítající řádově stovky GB až jednotky TB bez ohledu na velikost dostupné paměti počítače. Navržená metoda překonává ostatní dostupné nástroje v ohledech časových i paměťových požadavků a lze ji provozovat na běžných počítačích bez nutnosti investovat do speciálního HW vybavení. Extrahované n-gramy jsou využívány k obohacení modelů textových dokumentů a ke zvýšení přesnosti klasifikačních metod. Popis aplikace a jejího použití je k dispozici v zip archivu uloženém na katedře současně se zdrojovými kódy. Aplikace je připravena k použití. Byla vytvořena především proto, aby akceptovala vstup textových dokumentů ve franta-formátu a umožňovala reprezentovat textové dokumenty kromě samostatných slov i bigramy a itemsety, podobně jako v případě předchozí aktivity č. 21. Součástí je též nástroj Teraman (univerzální extraktor n-gramů), včetně zdrojových kódů a stručného manuálu.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat definované vstupy a poskytnout požadované výstupy, možnost nastavení nutných pro experimenty.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku. Výsledkem aktivity jsou publikace

• Tesař, R., Poesio, M., Strnad, V., Ježek, K.: Rozšíření bag-of-words modelu dokumentu: srovnání bigramů a 2-itemsetů. Publikováno na konferenci Znalosti 2007, Ostrava: VŠB - Technická univerzita, s. 131-142, ISBN 978-80-248-1279-3, únor 2007.

• Toman, M., Tesař, R., Ježek, K.: Vliv normalizace slov na klasifikaci textů. Publikováno na konferenci Znalosti 2007, Ostrava: VŠB - Technická univerzita, ISBN 978-80-248-1279-3, s. 360-363, únor 2007.
• Ježek, K., Hynek, J.: The Fight against Spam – A Machine Learning Approach. In: Proceedings of the 11th Int. Conf. on Electronic Publishing, ISBN 978-3-85437-292-9, pp.381-392 2007, Vienna, Austria, 2007.
• Češka, Z., Hanák, I., Tesař, R.: Teraman: A Tool for N-gram Extraction from Large Datasets. In: Proceedings of the IEEE 3rd International Conference on Intelligent Computer Communication and Processing (IEEE ICCP 2007), Cluj Napoca, Romania, pp. 209-216, ISBN 978-1-4244-1491-8, September 2007.

Číslo aktivity

23

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření modulu pro zpracování a vyhodnocení dat získaných z Internetu

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Součástí vytvářeného systému pro procházení webu a automatickou identifikaci internetových stránek podle tématu, o kterém pojednávají, bylo v první fázi projektu vytvoření aplikace označované jako webový spider. Jejím cílem bylo zadanou výchozí stránku zpracovat, získat z ní nebo určit předem definované údaje včetně odkazů na další stránky, které byly následně stejným způsobem zpracovány. Důležitou vlastností této aplikace je především její modularita, která jednoduchým způsobem dovoluje snadnou modifikaci. Možné využití tedy nepředstavuje jen automatické vytváření námětových korpusů, ale ve spojení s dalšími vhodnými moduly představuje vhodný prostředek pro ověření navržených algoritmů určených přímo k práci s webovým prostředím. Ve spojení s vhodným analyzátorem dat získaných z webového spidera je možné kompletně mapovat určitou tématickou doménu, v našem případě například servery obsahující závadné (protizákonné) materiály, které se vykytují v rámci určitého území – například České republiky nebo států Evropské unie. A právě tento analyzátor se v rámci této aktivity podařilo navrhnout a implementovat. Údaji poskytovanými tímto analyzátorem jsou statistiky počtu výskytů určitých slov na závadných webových stránkách, údaje o často se zde vyskytujících emailových adresách, analýza významnosti jednotlivých webových serverů z pohledu množství závadných dat apod. Popis vytvořeného modulu pro zpracování a vyhodnocení dat získaných z Internetu a jeho použití je k dispozici na katedře v zip archivu současně se zdrojovými kódy. Přiložen je i modul Webový Spider, jenž byl vytvořen v rámci aktivity, na kterou tato aktivita navázala. Aplikace je členěna na dva hlavní moduly, přičemž oba lze použít i samostatně. Původním záměrem bylo použít modul Webový Spider k získání požadovaných dat z internetu (primárně pornografie), které měly být plynule zpracovávány navazujícím modulem Webové Vazby analyzujícím získaná data. To se podařilo a z aplikace je možné získat poměrně zajímavá data, včetně zobrazení polohy serverů na mapě světa a další údaje. Oba moduly využívají společný zdroj, kterým je SQLite databáze.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - modulárně členěná, možnost různých nastavení.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, nastavení konfiguračních parametrů a testování běhu aplikace, zakomponování nově vytvořeného modulu, kontrolní úprava funkčnosti stávajících modulů.

Číslo aktivity

24

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Výběr a implementace algoritmů pro ohodnocení a výběr charakteristických položek

Zahájení aktivity

2.1.2007

Ukončení aktivity

30.12.2007

Popis aktivity

Pro obohacení modelů textových dokumentů bylo testováno použití n-gramů a itemsetů. Základním problémem ovšem bylo nejen jejich generování z textu, ale i výběr jen těch charakteristických položek, které nejvíce přispívají ke zvýšení úspěšnosti klasifikace. Po prostudování dostupné literatury a výběru jen těch přístupů, které obecně poskytují výborné výsledky, bylo přistoupeno k jejich implementaci s ohledem na zpracování co největšího množství textových dat. Popis aplikace a jejího použití je k dispozici v zip archivu uloženém na katedře současně se zdrojovými kódy. Zatím nelze stoprocentně zaručit, že generování funguje přesně tak jak má, navíc odkaz v článku na dlouhodobě existující a prověřený nástroj na internetu má daleko větší váhu (např. WEKA apod.).

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy - modulárně členěná, možnost různých relevantních nastavení.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola – ruční otestování aplikace, nastavení konfiguračních parametrů a testování poskytovaných výstupů. Výsledky získané díky dokončení této aktivity byly použity v publikaci

• Tesař, R., Poesio, M., Strnad, V., Ježek, K.: Rozšíření bag-of-words modelu dokumentu: srovnání bigramů a 2-itemsetů. In: Sborník referátů konference Znalosti 2007, Ostrava: VŠB - Technická univerzita, s. 131-142, ISBN 978-80-248-1279-3, únor 2007.

Číslo aktivity

25

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Lexikální databáze Verbalex obsahující valenční rámce českých sloves a jejich vazby na Princetonský WordNet v.2.0

Zahájení aktivity

1.7.2006

Ukončení aktivity

31.12.2009

Popis aktivity

Kompletování valenčních rámců českých sloves do synsetů, čištění významů a přiřazování překladových ekvivalentů z Princetonského WordNetu 2.0. Jde primárně o pracné manuální přiřazování vyžadující kvalifikované pracovníky. Do českého WordNetu byly doplněny informace o derivačních subsítích, které jej umožnily automaticky rozšířit o cca 29 000 literálů.

Skutečné Indikátory dosažení - výsledky aktivity

Zpracované rámce jsou k dispozici prostřednictvím webového rozhraní vytvořeného pro Verbalex. Rozšířený český WordNet obsahující derivační podsítě je přístupný skrze webový prohlížeč a editor DebVisdic.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky aktivity byly publikovány:

1. Pala, K., Horák, A., Rambousek, A., Vetulani, Z., Konieczka, P., Marciniak, J., Obrębski, T., Rzepecki, P., Walkowska, J.: DEB Platform tools for effective development of WordNets in application to PolNet, in: Proceedings of the LTC 2007 Conference (ed. by Zygmunt Vetulani et al),

2. Hlaváčková, D., Pala, K., Derivational Relations in Czech WordNet, in Proceedings of Balto-Slavonic Natural

Language Processing 2007, ACL Prague, 2007

Číslo aktivity

26

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Korpus syntaktických stromů včetně morfologické desambiguace

Zahájení aktivity

1.1.2007

Ukončení aktivity

30.6.2009

Popis aktivity

Pro účely automatického statistického vyhodnocení kvality vyvíjených algoritmů pro uspořádání výsledků syntaktické analýzy je zapotřebí připravit co nejrozsáhlejší korpus syntaktických frázových stromů systému synt. V roce 2007 jsme připravili základní korpus obsahující cca 5000 syntaktických stromů, který plánujeme dále rozšiřovat. Vstupní věty korpusu jsou založeny na pražském korpusu PDT-1.0, aby bylo možné syntaktickou informaci srovnávat s výsledky pražské školy, jejíž algoritmy i výstupy jsou založeny na jiných principech a algoritmech (složkové vs. závislostní stromy, stochastické vs. pravidlové systémy).

Skutečné Indikátory dosažení - výsledky aktivity

Korpus o rozsahu 5000 syntaktických stromů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Korpus dostupný prostřednictvím korpusového manažeru.

Číslo aktivity

27

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Korpus vzorových přepisů vět do konstrukcí TILu a přiřazení odpovídajících sémantických reprezentací

Zahájení aktivity

1.1.2007

Ukončení aktivity

31.12.2009

Popis aktivity

Pro účely logické analýzy vět v přirozeném jazyce jsme vytvořili testovací sadu vzorových (typových) přepisů českých vět do konstrukcí transparentní intenzionální logiky (TIL) spolu s přiřazením odpovídajících sémantických reprezentací. Uvedená sada obsahuje v současnosti jen základní kolekci vět pro testování inference v systému Dolphin. Tento systém je první prototyp implementující bázi znalostí TIL. Pro tvorbu konstrukcí se využívá valenční lexikon VerbaLex, který je také rozšiřován v rámci projektu.

Skutečné Indikátory dosažení - výsledky aktivity

Vytvořená sada vzorových přepisů českých vět do konstrukcí TILu využitá a testovaná v systému Dolphin.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vytvořený korpus je přístupný jako sada konstrukcí TILu ve formátu pro vstup systému Dolphin, což je zároveň formát, který poskytuje systém synt v případě automatické tvorby logických konstrukcí (tento systém v současnosti poskytuje tuto funkcionalitu jen pro velmi omezenou sadu vstupních slov).

Číslo aktivity

28

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Návrh a implementace guesseru - modulu pro automatické doplňování morfologické databáze češtiny

Zahájení aktivity

1.1.2007

Ukončení aktivity

31.12.2008

Popis aktivity

Modul automaticky doplňující nová slova do morfologické databáze češtiny je nutný pro realistický provoz ostatních komponent vytvářených v projektů - bude navržena jeho struktura a podle možnosti implementována. Vzhledem ke své obtížnosti aktivita nebude ukončena v r. 2007 a bude pokračovat v dalším období.

Skutečné Indikátory dosažení - výsledky aktivity

Současná podoba morfologickeho analyzátoru ajka neumožňuje snadné rozšíření o systematické zpracování derivační morfologie, stejně tak je obtížné až nemožné zachytit morfonologické alternace, a zjednodušit tak systém vzorů (kde aktuálně každá výjimka tvoří vlastní vzor). Zčásti je to dáno formátem dat, zčásti skutečností, že data jsou v podstatě interpretována v průběhu analýzy, takže zásadnější změny či rozšíření jejich formátu by znamenaly nutnost netriviálních zásahů do již tak rozsáhlého a komplikovaného kódu analyzátoru.

Byl proto vytvořen návrh morfologické analýzy, který umožní striktně oddělit data od vlastní analýzy, jež se omezí na pouhé vyhledání ve slovníku, konkrétně triviální průchod konečným automatem minimalizovaným algoritmy publikovanými J. Daciukem. Byly navrženy datové struktury, které tímto způsobem umožní realizovat nejen morfologickou analýzu a syntézu, ale i morfematickou analýzu, zjišťování odvozených slov a slova, ze kterého je analyzované slovo odvozené, a konečně odhad možné morfologické a morfematické analýzy u neznámých slov. Podstatnou vlastností je skutečnost, že složitost samotné analýzy (vyhledání ve slovníku) se nijak nemění při případných změnách či rozšiřování formátu dat vyvolaných reálnými potřebami aplikací a dále že formát dat a jejich správa nejsou nijak omezeny požadavky na časovou optimálnost, jako tomu bylo v dosavadním řešení, kdy data musela být do jisté míry uspořádána (strukturována) tak, aby analýza nad nimi byla co nejrychlejší.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Šmerk, Pavel. Morphemic Analysis: A Dictionary Lookup Instead of Real Analysis. In First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007. Brno: Masaryk University, 2007. od s. 77-85, ISBN 978-80-210-4471-5.

Číslo aktivity

29

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Detekce plagiátů (spamů) s využitím sémantických znalostí

Zahájení aktivity

1.1.2007

Ukončení aktivity

1.12.2009

Popis aktivity

Porovnání existujících koncepcí pro zjišťování plagiátů a příprava návrhu algoritmu, který bude schopen pracovat se znalostmi sémantické povahy. Aktivita bude pokračovat i v dalších letech.

Skutečné Indikátory dosažení - výsledky aktivity

V prvotních testech se ukázalo, že základem zjištění plagiátů je dobrá segmentace zdrojových textů.

Byl vytvořen systém pro automatickou segmentaci textu, založený na Latentní Sémantické Analýze (LSA). Výsledky byly shrnuty do dvou níže uvedených článků.

Byla zahájena práce na porovnání efektivity indexačních metod pro mnoharozměrné prostory, s cílem zjišťování

podobných dokumentů ve vektorovém prostoru u velký kolekcí textů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Řehůřek, Radim. On Dimensionality of Latent Semantic Indexing for Text Segmentation. Proceedings of the International Multiconference on Computer Science and Information Technology, Wisla, Poland, 2007, 2, od s. 347-356, 10 s. ISSN 1896-7094. 2007.

Řehůřek, Radim. Text Segmentation Using Context Overlap. Progress in Artificial Intelligence, GuimarĂŁes, Portugal : Springer Berlin / Heidelberg, 2007, 4874, od s. 647-658, 11 s. ISSN 0302-9743. 2007.

Číslo aktivity

30

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Rozpoznávání anaforických vztahů ve volných textech

Zahájení aktivity

1.7.2006

Ukončení aktivity

31.12.2009

Popis aktivity

Primárně bude věnována pozornost anaforickým vztahům pronominálním - budou testovány existující algoritmy pro rozpoznávání anaforických vztahů ve volném textu, posouzena jejich vhodnost pro češtinu. Budou se řešit vazby na moduly, které jsou pro rozpoznávání anaforických vztahů nezbytné, konkrétně na syntaktický analyzátor synt - v této souvislosti bude potřeba navrhnout vhodné formáty a notační konvence.

Skutečné Indikátory dosažení - výsledky aktivity

Pokračoval vývoj systému na automatickou analýzu anaforických vztahů, zejména s ohledem na potřebu použít tento systém pro data v různých formátech (zejména PDT 2.0, výstup ze syntaktického analyzátoru synt a BNC). Oproti předchozí versi systému se ukazuje jako výhodné jej rozčlenit do několika vrstev s různou úrovní abstrakce (zejména rozlišení markable úroveň vs. technická úroveň). Byly provedeny snahy adaptovat systém MARS vyvinutý na universitě ve Wolverhamptonu pro češtinu. Vzhledem k obtížnosti dané problematiky bude aktivita pokračovat i v r. 2008 a 2009.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Němčík, Václav. Enhancing Anaphora Resolution for Czech. In RASLAN 2007. 1. vyd. Brno : Masarykova Univerzita, 2007. od s. 57-62, 6 s. ISBN 978-80-210-4471-5.

Číslo aktivity

31

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Analýza problematiky vytváření grafiky a webovských presentací prostřednictvím dialogových systémů.

Zahájení aktivity

1.1.2007

Ukončení aktivity

31.12.2009

Popis aktivity

V oblasti analýzy vytváření webovských stránek a počítačové grafiky v kontextu sémantického webu aplikacemi v asistivních technologiích byla vytvořena základní taxonomie webovských presentací a navrženy odpovídající rámce pro formalizaci na bázi klasifikačních systémů. Byly započaty práce na vytváření grafických ontologií souvisejících se zpřístupňováním informací a možnosti vytváření internetových prezentací pro nevidomé. Byla vytvořena

základní testovací verze systému WebGen pro vytváření webovských presentací dialogovým způsobem a provedeny testy její základní funkčnosti, včetně požadavku přístupnosti vůči nevidomým (internetový standard Web Content Accessibility). Byla navržena metoda integrující popis grafického objektu do formátu SVG a technologie využití tohoto přístupu pro nevidomé uživatele. Rovněž byla vytvořena metoda popisu grafických scén na umožňující efektivní získávání informací o objektu dialogovým způsobem.

Skutečné Indikátory dosažení - výsledky aktivity

Publikace:

Kopeček, Ivan - Rajman, Martin. Project Internet for All - Creating Web Presentations and Graphics by means of a Dialogue System. In: Proceedings of the 2007 International Conference on Internet Computing ICOMP 2007. Las Vegas USA : CSREA Press, 2007. od s. 381-384, 4 s. ISBN 1-60132-044-2.

Bártek, Luděk - Kopeček, Ivan - Ošlejšek, Radek. Setting Layout in Dialogue Generating Web Pages. In: Text, Speech and Dialogue. 10th International Conference, Pilsen, Proceedings. Berlin : Springer, 2007. od s. 613-620, 8 s. ISBN 3-540-74627-7.

Implementace: Prototypová verze systému WegGen

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Viz příslušné sborníky

Číslo aktivity

32

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Vytvoření korpusu matematických textů pro vyhledávání na webu

Zahájení aktivity

1.1.2007

Ukončení aktivity

31.12.2008

Popis aktivity

Byl vytvořen korpus více než 100000 stran matematických textů časopisů Application of Mathematics, Archivum Mathematicum, Časopis pro pěstování matematiky, Časopis pro pěstování matematiky a fysiky Commentationes mathematicae Universitatis Carolinae, Czechoslovak Mathematical Journal a Mathematica Bohemica s využitím metadatového editoru editor.dml.cz projektu DML-CZ a na zkompletovaných datech (Czechoslovak Mathematical Journal a born-digital data Archivum Mathematicum) byl prováděn výzkum úspěsnosti metod strojového učení klasifikace matematických textů dle AMS 2000 classification scheme). Byly studovány možnosti indexování a vyhledávání strukturních informací (matematiky) ve formátech TeX a MathML.

Skutečné Indikátory dosažení - výsledky aktivity

Sojka, Petr - Řehůřek, Radim: Classification of Multilingual Mathematical Papers in DML-CZ. In: Proceedings of First Workshop of Recent Advances in Slavonic Natural Language

Processing RASLAN 2007. Brno : Masarykova univerzita, 2007.

od s. 89-96, 8 s. ISBN 978-80-210-4471-5.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

---

Číslo aktivity

33

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh nových metod automatického rozpoznávání dialogových aktů

Zahájení aktivity

15.1.2007

Ukončení aktivity

15.9.2007

Popis aktivity

Tato aktivita navazuje na aktivitu 2006-13 a jejím cílem je návrh nových metod automatického rozpoznávání dialogových aktů, které budou pracovat s větší přesností než metody existující. Převážná část existujících metod automatického rozpoznávání dialogových aktů modeluje věty pomocí jazykových modelů typu n-gram. Tyto modely modelují pouze lokální strukturu věty. Navrhované metody modelují globální větnou strukturu, k čemuž využívají pozici slov ve větě.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem této aktivity jsou tři nové metody automatického rozpoznávání dialogových aktů využívající globální pozici slov ve větě. První metoda, multiscale position, využívá popis věty na několika úrovních a vyhlazuje pravděpodobnostní odhady mezi těmito úrovněmi. Druhá metoda, non-linear merging, modeluje závislost mezi slovy a jejich pozicí ve větě pomocí neuronové sítě typu vícevrstvý perceptron. Třetí metoda, best position, využívá Bayesovský rámec a k odvození pravděpodobnosti dialogového aktu předpokládá nezávislost mezi slovem a jeho pozicí ve větě.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky nově navržených metod byly srovnány s výsledky referenční metody, unigram modelu. Přesnost všech navržených metod je vyšší, než přesnost referenční metody.

Metody byly publikovány v odborném časopise a v disertační práci:

Kral, P., Cerisasa, C., Kleckova, J.: Lexical Structure for Dialogue Act Recognition. In: Journal of Multimedia (JMM), ISSN : 1796-2048, Volume 2, Issue 3, June 2007, pp. 1-8.

Kral, P.: Automatic Recognition of Dialogue Acts. In: CoPhD. Thesis, Henri Poincaré University – Nancy 1 and University of West Bohemia in Pilsen, 2007.

Číslo aktivity

34

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Studie důležitosti prozódie v úloze automatického rozpoznávání dialogových aktů

Zahájení aktivity

20.9.2007

Ukončení aktivity

15.12.2007

Popis aktivity

Prozodickou informaci je možno použít pro rozpoznávání některých dialogových aktů. V angličtině je používána s uspokojivými výsledky zejména pro detekci otázek zjišťovacích, neúplných vět a souhlasů. V rámci této aktivity jsme se proto zaměřili na analýzu důležitosti prozódie pro rozpoznávání určitých dialogových aktů v českém jazyce. Množina dialogových aktů je následující: sdělení, otázka zjišťovací a ostatní otázky. Byly použity dva základní prozodické parametry: základní hlasivková frekvence (F0) a energie a srovnány dva klasifikátory: směs Gaussovských funkcí a vícevrstvý perceptron.

Skutečné Indikátory dosažení - výsledky aktivity

Bylo provedeno automatické rozpoznání výše zmíněných dialogových aktů pouze pomocí prozódie. Dále jsme provedli analýzu výsledků a rozbor prozodických atributů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky analýzy potvrzují, že prozódii je výhodné použít jako doplňkovou informaci pro rozpoznávání některých

dialogových aktů. Použijeme-li ji samostatně, výsledky rozpoznávání jsou nedostatečné (přesnost rozpoznávání přibližně jen 45%). Oba porovnávané klasifikátory (tj. směs Gaussovských funkcí a vícevrstvý perceptron) pracovaly se srovnatelnou přesností.
Výsledky experimentů byly publikovány na mezinárodní konferenci Speech and Computer (SPECOM 2007):
Kral, P., Cerisara, C., Kleckova, J.: Importance of Prosody for Dialogue Act Recognition. In: Proceedings of SPECOM'07, October 2007, Moscow, Russia.

Číslo aktivity
35
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Značkování korpusu předdefinovanými dialogovými akty
Zahájení aktivity
15.1.2007
Ukončení aktivity
31.10.2007
Popis aktivity
Ruční anotace korpusů dialogovými akty (i obecně) je velmi časově i finančně náročný proces. V rámci této aktivity jsme se proto zaměřili na návrh metody pro anotaci korpusu dialogovými akty, která by byla pokud možno co nejvíc automatická. Prostudovali jsme dostupné metody trénování založené na maximalizaci pravděpodobnosti. Většina těchto metody vychází z malého ručně anotovaného korpusu použitého pro natrénování základních modelů. Dále metody pracují iterativně; při každé iteraci vylepšují kvalitu modelů a zároveň rozšiřují korpus.
Skutečné Indikátory dosažení - výsledky aktivity
Byla navržena a implementována metoda trénování s učitelem i bez učitele, založená na algoritmu Expectation Maximization a míře důvěry. Prokázali jsme, že je možné tuto metodu úspěšně použít na poloautomatické značkování korpusu dialogovými akty.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Při použití metody automatické anotace dialogovými akty byl základní korpus, který se skládal z 1600 dialogových aktů, automaticky doplněn o dalších 500. Přesnost rozpoznávání se zlepšila přibližně o 10% (absolutně).
Výsledky této práce byly publikovány na mezinárodní IEEE konferenci International Conference on Acoustic, Speech and Signal Processing (ICASSP 2007):
Kral, P., Cerisasa, C., Kleckova, J.: Confidence Measures for Semi-automatic Labeling of Dialog Acts. In: Proceedings of the Conference ICASSP'07, Honolulu, Hawaii, USA, April 2007, pp. 153-156.

Číslo aktivity
36
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Návrh metody odhalující plagiáty textových dokumentů s využitím latentní sémantické analýzy
Zahájení aktivity
1.7.2007
Ukončení aktivity
31.12.2007
Popis aktivity
Dosud používané metody pro odhalování plagiátů hledají překryv textu mezi dvěma dokumenty bez uvažování sémantiky uvnitř dokumentů. Tato aktivita prozkoumává metody pro extrakci N-gramů různých velikostí, které jsou široce využívány v metodách odhalující plagiáty. N-gramy jsou základní stavební jednotkou v přirozeném

zpracování jazyka, které představují jednotlivé fráze uvnitř vět. Nad množinou frází je navrhována nová experimentální metoda, která využívá latentní sémantickou analýzu odhalující vztahy frází mezi ověřovanými dokumenty. Experimentální metoda pro odhalování plagiátů využívá latentní sémantickou analýzu, která je postavena na singulární dekompozici matic (SVD). Vstupní matice popisuje výskyty frází v jednotlivých dokumentech, kde sloupce reprezentují zkoumané dokumenty a řádky četnosti frází nacházející se v daném dokumentu. Fráze jsou extrahovány jako N-gramy o stanovené velikosti, pro jejichž získání byla provedena studie publikovaná na konferenci ITAT 2007. Vytvořená matice s výskyty jednotlivých frází je následně dekomponována na tři nezávislé matice a spočteny vztahy mezi frázemi a dokumenty. První matice obsahuje ortogonální vektory frází, které jsou pro naše experimentování nevýznamné. Druhá matice nese vlastní čísla udávající významnost jednotlivých dimenzí a třetí zahrnuje ortogonální vektory dokumentů. Právě ortogonální vektory dokumentů jsou pro nás nejdůležitější, protože po jejich normalizaci a přepočtu na korelační matici, udávají vzájemnou podobnost mezi dokumenty. Naše experimentální metoda tedy využívá singulární dekompozice k odhalení skrytého významu mezi plagiovanými dokumenty prostřednictvím společných frází.

Skutečné Indikátory dosažení - výsledky aktivity

Experimentální metoda využívající latentní sémantickou analýzu, která byla publikována na konferenci YRCAS 2007 a studie využití N-gramů pro odhalování plagiátů, publikovaná na konferenci ITAT 2007. Aktivita nenavazuje na žádnou předešlou aktivitu a do dalšího roku bude volně pokračovat aktivitou, která se bude zabývat implementací a ověřením navržené experimentální metody.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledkem aktivity jsou publikace:

• Češka, Z.: Využití N-gramů pro odhalování plagiátů. In: Proceedings of the ITAT 2007, Information Technologies - Applications and Theory, Polana, Slovakia, pp. 63-66, September 2007. ISBN 978-80-969184-6-1.

• Češka, Z.: The Future of Copy Detection Techniques. In: Proceedings of the 1st Young Researchers Conference on Applied Sciences (YRCAS 2007), pp. 5-10, Pilsen, Czech Republic, November 2007. ISBN 978-80-7043-574-8.

Řehůřek, R.: On Dimensionality of Latent Semantic Indexing for Text Segmentation, Proceedings of the 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA07)

Řehůřek, R.: Text Segmentation Using Context Overlap. Progress in Artificial Intelligence, Guimaraes, Portugal: Springer Berlin/Heidelberg, 2007, 4874, od s. 647-658, 11 s. ISSN 0302-9743. 2007.

Číslo aktivity

37

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Pořizování korpusu dotazů z reálného prostředí

Zahájení aktivity

1.6.2007

Ukončení aktivity

16.11.2007

Popis aktivity

Tato neplánovaná aktivita vznikla jako reakce na neplánovanou, ale velmi významnou příležitost získat reálná data z prostředí českých internetových vyhledávačů. V rámci řešení projektu se nám podařilo navázat spolupráci s firmou Seznam.cz a.s., která provozuje stejnojmenný internetový portál Seznam.cz. Tato firma nám poskytla data pocházející z reálného provozu jejich internetového vyhledávacího stroje. Celkové nám bylo poskytnuto cca 187 miliónů dotazů. Dotazy však bylo nutno zpracovat, jelikož dodané dotazy obsahovaly duplicitní dotazy, byly zakódované a hlavně většina z nich byla ve formě klíčových slov. Nás však pro zpracování sémantiky zajímaly pouze dotazy položené ve formě celých vět. Vzhledem k obrovskému množství dat bylo nutné, aby celé zpracování probíhalo automaticky. Nejprve bylo zapotřebí odfiltrovat duplicitní dotazy nepřímou hashovací tabulkou, poté bylo nutno dotazy dekódovat z URL kódování do textové podoby. Následně došlo k automatickému doplnění diakritiky navrženým algoritmem u dotazů, kde diakritika chyběla. V závěrečné fázi zpracování byl vytvořen statistický

dichotomický klasifikátor, který rozlišoval dotazy na dotazy položené ve formě klíčových slov a na dotazy položené ve formě celých vět. Celkově jsme tímto zpracováním získali 101165 vět. Tyto věty jsou velmi vhodné pro budoucí testování navrženého systému, jelikož se jedná o data z reálného provozu.

Skutečné Indikátory dosažení - výsledky aktivity

Automatickým dekódovacím a klasifikačním algoritmem bylo získáno 101165 vět ve formě celých vět z reálného prostředí internetového vyhledávače. Tyto věty budou sloužit zejména pro testování navržených algoritmů pro sémantickou analýzu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Věty byly manuálně zkontrolovány. Jejich kvalita a rozmanitost je velmi vhodná pro testování metod sémantické analýzy. Věty jsou uloženy v databázi na serveru a zálohovány na DVD.

Navržený systém byl prezentován na konferenci:

Konopík, M.: Analysis of User Queries to the Internet. In: PHDWS 2007 proceedings, Balatonfüred; Computer and Automation Research Institute, 2007, s.164-168, ISBN 978-963-311-365-3.

2.2.2. AKTIVITY NEUSKUTEČNĚNÉ v roce 2007

Číslo aktivity
Ke kterému dílčímu cíli se aktivita vztahuje
Název (cíl)aktivity
Zahájení aktivity
Ukončení aktivity
Popis aktivity
Důvody, proč se aktivitu nepodařilo uskutečnit

2.3.NÁKLADY PROJEKTU - 2007

2.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2007
Typ	skutečné
Organizace	Masarykova univerzita
Role organizace	

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1746**	**1472**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**210**	**210**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**79**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**35**	**35**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**128**	**79**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**200**	**0**
F9. CELKEM	**2396**	**1796**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

Rok	2007
Typ	skutečné
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**2948**	**2934**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**320**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**83**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**21**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**50**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**293**	**61**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**280**	**0**
F9. CELKEM	**3995**	**2995**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**97**	**97**

2.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2007
Typ	skutečné
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**4694**	**4406**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**530**	**210**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**79**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**118**	**35**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**21**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**50**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**421**	**140**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**480**	**0**
F9. CELKEM	**6391**	**4791**
	PŘEVOD DO fondu tis. Kč	**POUŽITÍ Z fondu tis. Kč**
F0. - Zúčtování s Fondem účelově určených prostředků	**97**	**97**

2.3.3. ZDŮVODNĚNÍ ZMĚN V ČERPÁNÍ

Masarykova univerzita:
Vzhledem k neočekávanému navýšení nákladů na některé plánované a vykonané služební cesty bylo potřeba přesunout částku 8 tis. Kč z provozních nákladů (F4) na cestovné (F7). Změna v rozdělení provozních nákladů (17 tis. Kč) mezi položky F3 a F4 byla způsobena aktuálními cenami výpočetní techniky, které se liší od cen platných při tvorbě návrhu.
Částky skutečně vynaložených nákladů v dalších položkách byly zachovány beze změn.

Západočeská univerzita:
Vzhledem k úspoře na dalších provozních nákladech a ke změnám v pracovních poměrech doktorandů (po obhájení disertace a splnění jimi realizovaných aktivit projektu přešli na jiná pracoviště) zůstalo nedočerpáno 75 tisíc Kč. Tyto prostředky jsou převáděny do příštího roku. Budou použity zejména k zapojování nadaných studentů do projektu formou dohod o provedení práce. Poněkud vyšší oproti předpokladu, avšak v rámci limitu, bylo čerpání výdajů na cestovné, způsobené zvýšením cen. Změna v čerpání prostředků na služby o 1 tisíc Kč je marginální a má stejné zdůvodnění.

2.3.4. NEVYUŽITÉ FINANČNÍ PROSTŘEDKY

Nevyužité finanční prostředky v částce 75 tisíc Kč byly převedeny do fondu účelově určených prostředků a budou využity v r. 2008 zejména na podporu zapojení nadaných studentů a zvýšené náklady na cestovné.

2.3.5. Seznam hmotného a nehmotného majetku pořízeného za sledované období

Pořadí	1
Název	Notebook VAIO s příslušenstvím
Podíl užití majetku pro řešení v %	46,7
Pořizovací cena v tis. Kč	123
Uznaný náklad v tis. Kč	57
Uhrazeno z dotace v tis. Kč	57
Datum dodání	27.7.2007
Datum zprovoznění	27.7.2007
Dodavatel	Autocont CZ a.s.

Pořadí	2
Název	Výpočetní a diskový server SUPERMICRO
Podíl užití majetku pro řešení v %	46,7
Pořizovací cena v tis. Kč	327
Uznaný náklad v tis. Kč	153
Uhrazeno z dotace v tis. Kč	153
Datum dodání	8.8.2007
Datum zprovoznění	9.8.2007
Dodavatel	M-Computers s.r.o. (orange&green)

Pořadí	3
Název	Diskový a výpočetní Server DELL PE2950
Podíl užití majetku pro řešení v %	100
Pořizovací cena v tis. Kč	223,6
Uznaný náklad v tis. Kč	223,6
Uhrazeno z dotace v tis. Kč	223,6
Datum dodání	14.11.2007
Datum zprovoznění	14.11.2007
Dodavatel	Axes Computers s.r.o., Kollárova 2116/1, 301 00 Plzeň

Pořadí	4
Název	2 kusy Notebook Latitude D430
Podíl užití majetku pro řešení v %	100
Pořizovací cena v tis. Kč	96,4
Uznaný náklad v tis. Kč	96,4
Uhrazeno z dotace v tis. Kč	96,4
Datum dodání	14.11.2007
Datum zprovoznění	14.11.2007
Dodavatel	Axes Computers s.r.o., Kollárova 2116/1, 301 00 Plzeň

3. ZÁMĚR A NÁVRHY PRO NÁSLEDUJÍCÍ OBDOBÍ - rok 2008

3.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

3.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce - koordinátor
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 8/
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

3.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Albrecht Štěpán Ing.** 810520/2061 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 496 377 632 402 albrs@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Brada Přemysl Ing. PhD. MSc.** 7007012111 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	3772435 brada@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Češka Zdeněk Ing.** 8207311244 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 zceska@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	31.25

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	30

Celé jméno, RČ	**Habernal Ivan Ing.** 830705/1764 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 habernal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Hanks Patrick Ph.D.** 400324 GB
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	hanks@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Hejtmánek Jan Ing.** 821101/2095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 hejtman2@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	30

Celé jméno, RČ	**Horák Aleš PhD.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Hynek Jiří ing. PhD.** 720506/2029 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632455 hynekj@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	40

Celé jméno, RČ	**Král Pavel Ing. PhD.** 760317/2049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pavelka Tomáš Ing.** 7909182083 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Ptáčková Helena** 705914/2079 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	5

Celé jméno, RČ	**Rohlík Ondřej Ing. PhD.** 7510031925 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632450 rohlik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Rychlý Pavel Mgr. PhD.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Sojka Petr RNDr. PhD.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	50

Celé jméno, RČ	**Steinberger Josef Ing. PhD.** 7909182127 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 479 jstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	25

Celé jméno, RČ	**Tesař Roman Ing.** 7909302379 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 romant@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	10

Celé jméno, RČ	**Toman Michal Ing.** 8007042054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 mtoman@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	60

Celé jméno, RČ	**Zíma Martin Ing. PhD.** 7405042073 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632431 zima@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

3.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2008

Pč.	Typ	Popis
1	návrhy změn v projektovém týmu a řešitelských týmech	Od 1.1.2008 byl přijat do řešitelského kolektivu O. Rohlík, který se vrátil z několikaletého působení na ETH Zürich. Jeho doménou budou práce na sémantickém webu. V budoucnu nahradí pracovníky, kteří ukončili či ukončí svá doktorská studia, v roce 2008 budou obhajovat disertace související s tématikou tohoto projektu a z univerzity buď již odešli nebo v nejbližší době odejdou do praxe. Dále se budou na řešení projektu částečně podílet (cca 30 %) interní doktorandi Ing. Ivan Habernal a Ing. Jan Hejtmánek, přijatí na katedru od 1.9.2007. Témata jejich doktorského studia a v budoucnu jejich disertačních prací úzce souvisejí s tématem řešeného projektu.

3.2. ČASOVÝ POSTUP PRACÍ - rok 2008

3.2.0. PŘEHLED DÍLČÍCH CÍLŮ PLÁNOVANÉ 2008

	Číslo	Dílčí cíl	Datum plnění
	1	Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím.	- 31.12.2007
	2	Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka.	- 31.12.2008
	3	Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce.	- 31.12.2009
	4	Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí.	- 31.12.2010

3.2.1. AKTIVITY PLÁNOVANÉ NA DALŠÍ OBDOBÍ - rok 2008

Číslo aktivity

01

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Sémantické anotování korpusu a vytváření anotačních schémat - dokončení

Zahájení aktivity

2.1.2008

Ukončení aktivity

1.6.2008

Popis aktivity

V rámci této aktivity budou vytvořena anotační schémata pro zbylá témata, která nebyla pokryta v aktivitě 2007-04. Vytvořená anotační schémata se následně použijí pro pokračování prací na sémantickém anotování korpusu (aktivita 2007-05). V rámci aktivity 2007-05 bylo vytvořeno dostatečné množství trénovacích dat nutných pro vývoj algoritmů sémantické analýzy (aktivity 2008-02 a 2008-03), avšak pořízení většího množství trénovacích dat umožní další zlepšování algoritmů sémantické analýzy. Při této aktivitě bude zpracováván zejména korpus typických dotazů (získaný v aktivitě 2006-04) a částečně bude zpracován i korpus z aktivity 2007-37.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem aktivity bude sémanticky anotovaný korpus a sada anotačních schémat. Měřitelným parametrem bude počet anotovaných vět z korpusu (2006-04) a dosažená mezianotátorská shoda.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Anotovaný korpus bude uložen do databáze projektu. Ověřit počet anotovaných vět a mezianotátoskou shodu bude možné dodaným algoritmem.

Číslo aktivity

02

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Sémantická analýza lexikálních tříd

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

Cílem této aktivity je navázat na aktivitu 2007-06 - Identifikace lexikálních tříd. Daná množina lexikálních tříd bude dále sémanticky zpracovávána a bude vytvořen algoritmus formalizace lexikálních tříd. Extrakce významu lexikální třídy bude využívat znalostní přístup založený na syntaktické analýze a následném vyhodnocení/překladu. Výsledky této aktivity budou sloužit jako základní stavební pilíře sémantické analýzy (aktivita 2008-03).

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem aktivity bude sémantická reprezentace lexikální třídy specifickým formalismem. Kontrola výsledků bude provedena testováním algoritmů na množině vět získaných v aktivitě 2007-37 a měřením úspěšnosti správné extrakce sémantiky dané lexikální třídy.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky budou publikovány v odborném periodiku. Výsledný algoritmus bude k dispozici ke stažení na webových stránkách projektu.

Číslo aktivity

03

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj základního algoritmu pro automatickou sémantickou analýzu dotazů

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Cílem této aktivity je vytvořit základní algoritmus pro sémantickou analýzu dotazů. Součástí aktivity bude zvolení vhodného stochastického modelu pro zpracování sémantiky, vytvoření algoritmu pro trénování zvoleného modelu a vytvoření algoritmu pro vlastní sémantickou analýzu na základě natrénovaného modelu. Pro trénování se bude využívat anotovaný sémantický korpus (vytvořený v rámci aktivit 2007-05 a 2008-02). Tvorba této aktivity úzce souvisí s naplněním aktivity 2008-02, jelikož lexikální třídy budou tvořit základ pro sémantickou analýzu (používáme hybridní pravidlový a stochastický přístup). Výsledkem této aktivity bude tzv. baseline (základní) verze algoritmu pro sémantickou analýzu. Tento algoritmus bude v dalších fázích zpracování projektu vylepšován a zdokonalován.

Plánované indikátory dosažení - očekávané výsledky aktivity

Předpokládaným výsledkem bude úspěšnost algoritmu v procentech shody s ručně vytvořenými sémantickými stromy. Algoritmus bude testován na datech, která nebudou použita pro trénovaní algoritmu. Robustnost algoritmu bude ověřena na větách získaných v rámci aktivity 2007-37.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky algoritmu budou publikovány v odborné literatuře. Úspěšnost bude možno ověřit na vytvořených datech srovnáním výsledků algoritmu s výsledky získanými lidskými anotátory.

Číslo aktivity

04

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Trénování akustických modelů

Zahájení aktivity

1.12.2007

Ukončení aktivity

20.12.2008

Popis aktivity

V současné době jsou k dispozici natrénované akustické modely založené na umělých neuronových sítích nebo na směsích Gaussovských funkcí pro dva korpusy (ovládání šachové hry a dotazy na vlaková spojení), které má řešitelský kolektiv k dispozici z doby před zahájením práce na projektu COT-SEWing. Výsledkem nahrávání korpusů v rámci projektu je pořízení většího množství řečových dat, která budou moci být využita pro natrénování nových akustických modelů. Předpokládá se, že větší množství trénovacích dat povede k zvýšení úspěšnosti rozpoznávání. Trénování a testování se zaměří především na kontextově závislé modely.

Plánované indikátory dosažení - očekávané výsledky aktivity

Natrénované akustické modely pro automatický rozpoznávač řeči JLASER.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testy úspěšnosti rozpoznávání souvisle vyslovených vět.

Číslo aktivity

05

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj rozpoznávače JLASER

Zahájení aktivity

4.1.2008

Ukončení aktivity

20.12.2008

Popis aktivity

Automatický rozpoznávač řeči JLASER je nyní ve verzi 1.1, předpokládá se, že vývoj bude nadále pokračovat podle potřeb projektu.

Plánované indikátory dosažení - očekávané výsledky aktivity

K dispozici budou dány zdrojové kódy vytvořených programů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Testy úspěšnosti rozpoznávání vyslovených souvislých vět.

Číslo aktivity

06

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření mapy slovních kategorií pro pořízené korpusy a návrh vhodného „neuronového“ systému pro kategorizaci dokumentů

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Cílem aktivity bude dokončení testů a natrénování mapy slovních kategorií založené na Kohonenově samoorganizující mapě. Cílem testů bude ověřit chování mapy v případě rozsáhlých korpusů, porovnat vytvořené slovní kategorie pro různé modifikace vstupů, které vyjadřující sémantický kontext dokumentů, a ověřit různé algoritmy trénování. Kohonenovu mapu je v zásadě možné trénovat dvěma způsoby, a to sekvenčním algoritmem, který je však pro rozsáhlé korpusy značně časově náročný, nebo tzv. „batch“ algoritmem, jehož rychlost je mnohem vyšší. Dosavadní testy využívaly převážně sekvenční algoritmus. V další části se zaměříme na implementaci „batch“ trénovacího algoritmu do stávajícího systému a testování jeho vlastností při vytváření mapy slovních kategorií. Dalším úkolem bude návrh systému pro kategorizaci dokumentů na základě vytvořených slovních kategorií. V systému WEBSOM bude kategorizace dokumentů prováděna opět Kohonenovou mapou (tzv. mapou dokumentů), kde vstupem této mapy bude předzpracovaný vektor slovních kategorií obsažených v kategorizovaném dokumentu. V rámci aktivity bude vytvořena a otestována mapa dokumentů a bude provedeno optimální nastavení parametrů této mapy s ohledem na výsledky kategorizace dokumentů. Dále bude ověřeno, zda je možné (a za jakých podmínek) nahradit mapu dokumentů jinou neuronovou sítí. Doposud provedené testy ukazují, že by Kohonenova mapa mohla být nahrazena jinou sítí učenou bez učitele, např. sítí ART-2.

Plánované indikátory dosažení - očekávané výsledky aktivity

Implementace „batch“ algoritmu, trénování a jeho začlenění do stávajícího systému, modifikace vstupů neuronové sítě, finální natrénování mapy slovních kategorií. Vytvoření a natrénování mapy dokumentů a ověření, které z neuronových sítí by bylo možné využít pro kategorizaci dokumentů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Statistika úspěšnosti kategorizace dokumentů pro testované neuronové sítě.

Číslo aktivity

07

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Sumarizace textů založená na tenzorové LSA

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.12.2007

Popis aktivity

Myšlenky single-document sumarizátoru založeného na latentní sémantické analýze byly v předchozí etapě projektu aplikovány na multi-document sumarizaci. Témata LSA se zde tvoří na základě společného výskytu termů v větách. Navržená metoda však nebere v úvahu společný výskyt termů v dokumentech. Hodláme tedy přidat třetí „dokumentovou“ dimenzi do vstupních dat LSA. Tenzorová LSA metoda (TLSA) by měla zpřesnit vzniklá témata. Dále bude testována nová metoda tvorby souhrnu z témat LSA – MMR (Maximal Marginal Relevance). Myšlenkou MMR je, že věta vkládaná do souhrnu by měla být co nejvíce podobná dotazu uživatele a co nejvíce rozdílná od vět, které již jsou v souhrnu obsaženy.

Plánované indikátory dosažení - očekávané výsledky aktivity

Multi-document sumarizační systém založený na TLSA.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Experimenty na DUC korpusech a kolekci českých textů. Porovnání s metodou založenou na maticové LSA (viz předchozí etapa projektu).

Číslo aktivity

08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Řazení vět v souhrnu

Zahájení aktivity

1.2.2008

Ukončení aktivity

30.9.2008

Popis aktivity

Při přechodu od single-document sumarizace k multi-document sumarizaci se objevil nový problém. Věty pocházející z různých dokumentů musejí být v souhrnu seřazeny. Cílem proto bude navržení metody, která seřadí věty na základě výskytů jednotlivých entit textu. Myšlenkou je, aby věty, které se vyskytnou ve výsledném souhrnu vedle sebe, obsahovaly co nejvíce výskytů společných entit.

Plánované indikátory dosažení - očekávané výsledky aktivity

Modul, který seřadí věty souhrnů v XML kolekci.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Systém bude porovnán se systémem vytvořeným anotátory. Bude tedy nutné, aby nejprve anotátoři vytvořili referenční („správné“) pořadí vět, které bude potom porovnáno s tím, které vytvoří automatický systém.

Číslo aktivity

09

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Online vyhledávací a sumarizační systém

Zahájení aktivity

1.2.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Cílem této aktivity bude nasazení sumarizačního systému online a přiblížení se pojmu sémantický web. Systém bude pracovat následovně: Uživatel vloží dotaz, který by měl být dostatečně bohatý, aby vymezil dané téma. Systém následně vyhledá nejrelevantnější dokumenty k vloženému dotazu. Pak bude spuštěn řetěz sumarizačních modulů, které jsou paralelně vyvíjeny (převod do XML včetně tokenizace textu na věty, označení vedlejších vět, označení jmenných entit, sumarizátor vytvářející extrakt, modul pro kompresi vět, modul pro seřazení vět, modul pro korekci anaforických referencí). Výsledný souhrn dokumentů bude vrácen uživateli jako odpověď systému na uživatelův dotaz.

Plánované indikátory dosažení - očekávané výsledky aktivity

Online vyhledávací a sumarizační systém veřejně přístupný z webových stránek projektu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Systém bude testován několika uživateli. Vždy bude zaznamenán jejich dotaz a odpověď systému uživateli. Uživatel vyjádří svou spokojenost/nespokojenost s odpovědí vyplněním online dotazníku.

Číslo aktivity

10

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Experimenty v systému pro vyhledávání

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

V rámci projektu navrhujeme prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jádrem vyhledávání je thesaurus EuroWordNet a sumarizátor je založen na latentní sémantické analýze. Současné řešení obsahuje především možnosti zpracování anglického a českého jazyka. V rámci aktivit předchozího roku jsme provedli implementaci systému pro rozšiřování dotazů a disambiguaci (aktivity 2007-17 a 2007-18). Nyní plánujeme provádět testy na celém prototypovém systému.

Plánované indikátory dosažení - očekávané výsledky aktivity

Navržený systém bude testován z hlediska relevance a úplnosti vyhledávaných dokumentů při aplikaci jednotlivých modulů. Bude zajištěna a testována především součinnost a vliv na obdržené výsledky.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Systém bude testován na relevanci získávaných výsledků jak pro české a anglické prostředí, tak i při křížovém zpracování. Provedeno bude také jako v předešlých aktivitách srovnání výsledků s přístupem aplikovaným ve vyhledávači Google.

Číslo aktivity

11

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity
Extrakce informace z webových sídel
Zahájení aktivity
2.1.2008
Ukončení aktivity
30.9.2008
Popis aktivity
Prostředí webu se postupně vyvinulo v obecný zdroj informací uchovávaných převážně v částečně strukturovaném formátu HTML. Stávající Web obsahuje data, která jsou určena pro prohlížení uživatelem, jenž zobrazenému textu přiřadí správnou sémantickou informaci. Jednotlivé zdroje vyjadřují informace v rozdílných formátech a různým způsobem, což pro člověka nepředstavuje větší problém, ale komplikuje jejich porozumění počítačem. Taková situace skýtá velké množství netriviálních úloh, které ve výsledku mohou vést k transformaci stávajících zdrojů do tzv. sémantického webu. Jednou větví výzkumu v této oblasti je extrakce informací z dat (IE; Information Extraction). Cílem extrakce dat je v našem případe získat z webových stránek čistý text a přidružená metadata, např: čas publikování příspěvku, autora, kategorii, název, perex. Vlastní text článku je však v praxi velmi často nesouvislý a je přerušen multimediálními daty, případně reklamou, což představuje komplikace.
Plánované indikátory dosažení - očekávané výsledky aktivity
Navržený systém bude umožňovat extrakci vybraných dat z webových stránek.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Systém bude testován na přesnost a úplnost extrakce pro české a anglické prostředí. Porovnání extraktů bude provedeno ručně zaškolenými pracovníky.

Číslo aktivity
12
Ke kterému dílčímu cíli se aktivita vztahuje
1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...
Název (cíl)aktivity
Příprava a vytvoření datové kolekce pro testování dotazů nad sémantickým webem
Zahájení aktivity
2.1.2008
Ukončení aktivity
30.6.2008
Popis aktivity
Navrhovaná datová kolekce bude obsahovat informace o studijních oborech softwarového inženýrství. Potřebné informace budou čerpány z informačních systémů technických vysokých škol a univerzit v ČR. Jako alternativa bude dále navržena datová kolekce katastrof, kde požadované informace budou získány z internetového serveru www.katastrofy.com. Na základě vytvořené datové kolekce bude navržena a vytvořena ontologie v jazycích RDF, RDFS a OWL. Při návrhu ontologie bude kladen důraz na různé závislosti, které lze vyjádřit logickými pravidly.
Plánované indikátory dosažení - očekávané výsledky aktivity
Popisované datové kolekce a odpovídající ontologie v jazycích RDF, RDFS a OWL.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Využití vytvořených kolekcí k otestování navržených formalismů a algoritmů navržených v aktivitě „Transformace ontologie na logický pravidlově orientovaný program a zpět“ (aktivita 2008-13).

Číslo aktivity
13
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity

Transformace ontologie na logický pravidlově orientovaný program a zpět

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Návrh formalismů, které umožní transformovat ontologii zapsanou v jazycích RDF, RDFS a OWL na logický program, který bude založen na pravidlech. Ze zápisu libovolné ontologie je patrné, že mezi jednotlivými částmi ontologie jsou definovány různé vztahy, které lze převést na zápis logického pravidla. Také logický program, který bude vyhovovat jistým podmínkám, bude možné transformovat do tvaru ontologie.

Plánované indikátory dosažení - očekávané výsledky aktivity

Návrh a realizace formalismů a algoritmů, které umožní transformaci ontologie na logický pravidlově orientovaný program. Specifikace podmínek, kdy je možné provést transformaci logického programu na odpovídající ontologii.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Na základě položeného dotazu bude dokázáno, že ontologie a odpovídající logický program poskytují shodnou odpověď, tj. že daná ontologie a logický program jsou ekvivalentní.

Číslo aktivity

14

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Editor pro anotaci korpusu dialogovými akty

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

Aktuální korpus anotovaný dialogovými akty je příliš malý pro další experimenty a obsahuje pouze několik základních dialogových aktů. Cílem této aktivity, navazující na 2006-14, bude vytvořit nástroj, který by umožnil anotaci korpusu dialogovými akty. Důraz zde bude kladen na rychlost anotace, čemuž bude odpovídat uživatelské rozhranní vyvíjeného nástroje, a na obecnost algoritmu anotace (množinu dialogových aktů si definuje uživatel).

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace, pomocí níž bude možné anotovat korpusy dialogovými akty.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Manuální ověření funkčnosti při anotaci korpusu.

Číslo aktivity

15

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Automatické rozpoznávání dialogových aktů

Zahájení aktivity

1.4.2008

Ukončení aktivity

15.12.2008

Popis aktivity

Tato aktivita navazuje na 2006-13 a 2007-34. V rámci těchto předchozích aktivit byly využity k rozpoznávání

dialogových aktů dva hlavní zdroje informací: lexikální informace a prozódie. Některé studie ukazují, že je výhodné tyto informace doplnit o tzv. dialogovou historii (časovou sekvenci po sobě jdoucích dialogových aktů). Cílem této aktivity je zaměřit se na tuto v našich metodách zatím chybějících informaci, prostudovat dostupné metody, které tuto informaci využívají, a vybrat z nich nejvhodnější. Následně analyzovat přínos vybrané metody a na základě této analýzy rozhodnout o integraci do stávajícího systému.

Plánované indikátory dosažení - očekávané výsledky aktivity

Studie dostupných metod automatického rozpoznávání dialogových aktů využívajících dialogovou historii. Výběr optimální metody.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Experimentální ověření vybrané metody na českém korpusu dialogových aktů.

Číslo aktivity

16

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Tvorba českého korpusu plagiátů textových dokumentů

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

K dispozici dosud není žádný český korpus, který by mohl být použit k ověření metod odhalující plagiáty textových dokumentů. Proto v rámci této aktivity bude manuálně vytvořen korpus zahrnující textové dokumenty, které jsou s různým procentuelním podílem okopírovány z jiných zdrojů, včetně kombinací několika různých zdrojů. Každý dokument bude označkován pro následné porovnávání a uložen v podobě XML souboru.

Plánované indikátory dosažení - očekávané výsledky aktivity

Český korpus plagiátů textových dokumentů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Použití pro otestování různých metod odhalující plagiáty textových dokumentů.

Číslo aktivity

17

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Návrh a vývoj úložiště textových dokumentů

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Pro uchování korpusu plagiátů s možností centralizovaného přístupu bude navrženo datové úložiště. Spolu s úložištěm bude vytvořena aplikace pro import textových dokumentů a dokumentů ve specializovaném formátu, jenž je používán korpusem ČTK. Tato aplikace usnadní hromadné naplnění databáze experimentálními daty a usnadní testování metod odhalující plagiáty. Import bude zaměřen na vložení korpusu pořízeného v rámci aktivity 2008-16 a dalších textových dokumentů pro výzkumné potřeby.

Plánované indikátory dosažení - očekávané výsledky aktivity

Centralizované úložiště textových dokumentů spolu s importem dat.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Použití při testování metody odhalující plagiáty plánované v rámci aktivity 2008-17 a u dalších baseline metod.

Číslo aktivity
18
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Implementace a experimentální ověření metody odhalující plagiáty textových dokumentů s využitím latentní sémantické analýzy
Zahájení aktivity
2.1.2008
Ukončení aktivity
31.12.2008
Popis aktivity
Z důvodu dosud nízkého zájmů o vyhledávání plagiátů v České republice a ověřování původu studentských prací bude implementována metoda s optimalizací pro české prostředí. Do výběru bude též zahrnuta podpora pro anglický jazyk. Tato aktivita volně navazuje na předchozí aktivitu 2007-36, která se zabývala návrhem experimentální metody pro identifikaci plagiátů textových dokumentů s využitím latentní sémantické analýzy. Součástí této aktivity bude též ověření experimentální metody a ověření její funkčnosti oproti jiným již existujícím metodám.
Plánované indikátory dosažení - očekávané výsledky aktivity
Implementace specializované metody využívající latentní sémantickou analýzu pro odhalování plagiátů textových dokumentů.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Navržená metoda bude testována na relevanci výsledků s označkovaným českým korpusem vytvořeným v rámci aktivit roku 2008.

Číslo aktivity
19
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Optimalizace a finální implementace metody Teraman
Zahájení aktivity
2.1.2008
Ukončení aktivity
30.9.2008
Popis aktivity
Tato aktivita volně navazuje na aktivitu 2007-22, v jejímž rámci byl vytvořen nástroj Teraman pro extrakci N-gramů z rozsáhlých textových dat. U navržené metody plánujeme optimalizovat modul odstraňující duplikáty N-gramů, které vznikají dělením vstupního korpusu na menší části při nedostatku paměti. Tato optimalizace by měla zapříčinit podstatné snížení časových požadavků u extrémně velkých korpusů čítající desítky GB a více.
Plánované indikátory dosažení - očekávané výsledky aktivity
Optimalizovaná metoda Teraman s nižšími časovými požadavky pro extrémně velké textové korpusy.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Porovnání času potřebného pro extrakci N-gramů u korpusů čítající desítky GB a více s dříve navrženou metodou.

Číslo aktivity

20

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

SPOT - online slovník odborné terminologie - další vývoj a rozšiřování obsahu

Zahájení aktivity

3.1.2007

Ukončení aktivity

31.12.2008

Popis aktivity

Slovník odborné terminologie bude sloužit dvěma účelům: jako zdroj referenčních překladů a jako platforma pro diskuse při jejich „ustalování“. SPOT bude rovněž možné nasadit v dalších aplikačních oblastech jako je sémantický web, multilinguální rozhraní, rozšiřování dotazů v rámci IR, thesaury apod. Další rozvoj slovníku bude zajištěn probíhajícím vývojem. Hlavní úkoly: zvýšení uživatelského komfortu úpravami uživatelského rozhraní (kontextová nápověda k vybraným funkcím, doplnění nápovědy v angličtině a němčině, zobrazování historie editačních akcí, filtrování podle stavů slov); přepracování GUI pro výsledky hledání ve slovníku; doplnění o další důvěryhodné zdroje pro zobrazování nalezených termínů v kontextu; ochrana slovníkové databáze proti zcizení obsahu; dokončení lokalizace projektu do němčiny; doplnění funkcí pro dávkovou aktualizaci slovní zásoby ze souborů s korpusy, včetně kontrol a odstraňování duplicit; zavedení a/č korpusu z oboru ICT (řádově desítky tisíc překladových dvojic); zavedení n/č korpusu z oboru technika; vytvoření kódu pro portlet SPOT pro účely šíření na další portály; doplnění funkcí pro vkládání obrázků k termínům. Pokračování provozu (aktivní tvorba obsahu) blogu pro uživatele na adrese http://www.blogspot.cz . Další podrobnosti jsou dostupné z wiki stránek projektu: http://wiki.kiv.zcu.cz/SlovnikTerminologie/HomePage .

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční aplikace schopná akceptovat požadované vstupy a poskytnout požadované výstupy.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Číslo aktivity

25

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Formalizace lexikální databáze Verbalex obsahující valenční rámce českých sloves a jejich vazby na princetonský WordNet v.2.0 a v.3.0.

Zahájení aktivity

1.7.2007

Ukončení aktivity

Popis aktivity

Pro komplexní valenční rámce získané při budování databáze Verbalex budou hledány vhodné notační varianty, které budou použitelné jak v syntaktickém analyzátoru, tak i v aplikaci pro manipulaci se sémantickými rámci a reprezentacemi. V souvislosti s pracemi na světovém wordnetovém gridu budou studovány použitelné datové struktury a jejich reprezentace s ohledem na vazby s princetonským (a dalšími) Wordnetem. Předmětem pozornosti budou též vhodné nástroje pro světový wordnetový grid.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem budou soubory rámců použitelných v experimentech s analyzátorem Synt a při experimentech s přístupy k webu v přir. jazyce.

U vazeb na princetonský Wordnet půjde o získání dostatečného seznamu translačních ekvivalentů napojených na

ILI.
Nástroje pro světový wordnetový grid.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Výsledky budou prezentovány ve formě publikací.

Číslo aktivity
26
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Tvorba modelu české syntaxe na základě korpusu syntaktických stromů
Zahájení aktivity

Ukončení aktivity

Popis aktivity
Vytvořený korpus syntaktických stromů bude použit pro vývoj a testování algoritmů na modelování české syntaxe. Současně bude tento korpus dále rozšířen.
Plánované indikátory dosažení - očekávané výsledky aktivity
zkvalitněný model české syntaxe použitý v syntaktickém analyzátoru synt.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
rozšířený korpus syntaktických stromů
publikace

Číslo aktivity
27
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Návrh formalismu pro práci s konstrukcemi TILu jako sémantické reprezentace českých vět
Zahájení aktivity

Ukončení aktivity

Popis aktivity
Pro zkvalitnění logické analýzy věty v přirozeném jazyce je zapotřebí podrobný návrh formalismu pro tvorbu konstrukcí a typování vstupních slov. Formalismus pro tvorbu konstrukcí bude založen (v souladu s principem kompozicionality) na pravidlech syntaktického analyzátoru. Typování vstupních slov bude sledovat zvolenou ontologii - hypero/hyponymickou hierarchii z princetonského WordNetu.
Plánované indikátory dosažení - očekávané výsledky aktivity
Rozšířený popis formalismu tvorby konstrukcí TIL
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
publikace

Číslo aktivity
28
Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a implementace guesseru - modulu pro automatické doplňování morfologické databáze češtiny

Zahájení aktivity

1.1.2008

Ukončení aktivity

Popis aktivity

V rámci aktivity bude vytvořen návrh algoritmu guesseru a posléze i zkušební implementace. Guesser je velmi potřebným nástroj pro rozšiřování a doplňování morfologické databáze češtiny zejména s ohledem na přistupování k webu a s ohledem na práci s terminologiemi.

Plánované indikátory dosažení - očekávané výsledky aktivity

základní verze guesseru

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

testování základní verze, publikace

Číslo aktivity

29

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Detekce plagiátů (spamů) s využitím sémantických znalostí

Zahájení aktivity

1.1.2007

Ukončení aktivity

Popis aktivity

V návaznosti na předchozí aktivitu rozpracování algoritmu (ev. i nástroje) pro detekci plagiátů využívajícího sémantické znalosti (využití Wordnetu, příadně databáze Verbalex).

Plánované indikátory dosažení - očekávané výsledky aktivity

ověření detekčního algoritmu

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace

Číslo aktivity

30

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Rozpoznávání anaforických vztahů ve volných textech

Zahájení aktivity

1.1.2007

Ukončení aktivity

Popis aktivity

Pokračování prací na modulu pro rozpoznávání anaforických vztahů (primárně zájmenných), vazby na syntaktický analyzátor Synt a český Wordnet. Experimenty s použitím komplexních valenčních rámců z databáze Verbalex.

Plánované indikátory dosažení - očekávané výsledky aktivity

Pokročilejší verze modulu pro rozpoznávání anaforických vztahů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
publikace

Číslo aktivity
31
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Návrh a vývoj algoritmů pro vytváření grafiky a webovských presentací prostřednictvím dialogových systémů.
Zahájení aktivity
1.1.2007
Ukončení aktivity

Popis aktivity
Budou pokračovat práce na vytváření grafických ontologií souvisejících se zpřístupňováním informací a možnosti vytváření internetových prezentací pro nevidomé a bude provedena analýza využitelnosti technologií sémantického webu pro vyhledávání grafických objektů na webu a pro pozice grafických objektů. Bude pokračovat testování systému WebGen pro vytváření webovských presentací dialogovým způsobem s cílem zefektivnění použitých dialogových strategií a budou provněž provedeny testy této verze nevidomými uživateli včetně ověření požadavku přístupnosti (internetový standard Web Content Accessibility). Bude vytvořena první verze systému WebGen. Bude pokračovat práce na metodách integrujících popis grafického objektu do formátu SVG a bude testována technologie využití tohoto přístupu pro nevidomé uživatele. Rovněž bude testována metoda strukturálního popisu grafických scén.
Plánované indikátory dosažení - očekávané výsledky aktivity
První verze systému WebGen.
Jednotlivé metody.
Testy s nevidomými uživateli a jejich vyhodnocení.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
publikace

Číslo aktivity
32
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Klasifikace matematických textů ve vytvořeném korpusu
Zahájení aktivity

Ukončení aktivity

Popis aktivity
Indexace dostupných matematických textů v nástroji manatee, návrh tokenizace, dotazových vzorů pro hledání citací v textu a rozhraní pro dotazování matematiky (nejlépe přizpůsobením nástroje bonito2 pro matematiku). Po shromáždění co nejreprezentativnější kolekce matematických článků naučit co nejpřesnější klasifikátor matematických textů dle AMS 2000 classification scheme a oklasifikovat dosud neklasifikované články. Na to naváže plánovaná aktivita v roce 2008 a 2009 zahrnující klasifikaci a strojové učení klasifikace matematických textů dle AMS 2000 classification scheme.
Plánované indikátory dosažení - očekávané výsledky aktivity
trénovaný klasifikátor

klasifikace matematických textů

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace

3.2.2. NÁVRH ZMĚN V ŘEŠENÍ PROJEKTU - rok 2008

Pč.	Typ	Popis
1	návrh změn v řešení projektu	K dnešnímu datu žádné změny v řešení projektu plánovány nejsou. Ovšem realita může být jiná - v uplynulém roce bylo nutno vyřešit pět dalších aktivit (33 - 37), aby aktivity plánované mohly být naplněny. Předpokládáme, že stejně tak tomu bude i v roce 2008.

3.3. NÁKLADY PROJEKTU - rok 2008

3.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2008
Typ	požadované
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**2965**	**2945**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**100**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**30**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**100**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**450**	**0**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**350**	**0**
F9. CELKEM	**3995**	**2995**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**75**	**75**

Rok	2008
Typ	požadované
Organizace	Masarykova univerzita
Role organizace	spolupříjemce

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1822**	**1521**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**60**	**40**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**60**	**40**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**180**	**121**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**200**	**0**
F9. CELKEM	**2322**	**1722**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

3.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2008
Typ	požadované
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**4787**	**4466**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**60**	**40**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**160**	**40**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**30**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**100**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**630**	**121**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**550**	**0**
F9. CELKEM	**6317**	**4717**
	PŘEVOD DO fondu tis. Kč	**POUŽITÍ Z fondu tis. Kč**
F0. - Zúčtování s Fondem účelově určených prostředků	**75**	**75**

3.3.3. NÁVRH ZMĚN V NÁKLADECH - rok 2008

Pč.	Typ	Popis
1	návrh změn v nákladech	Kromě čerpání nevyužitých finančních prostředků z roku 2007 (převedených do fondu účelově určených prostředků) na dohody o pracích (zejména pro studenty spolupracující na tvorbě korpusů a programovém řešení úloh) a podporu aktivit nově přijatých doktorandů nejsou žádné další změny v nákladech na řešení projektu plánovány.

4. PŘÍLOHY

4.1. ZPRÁVA O POSTUPU ŘEŠENÍ PROJEKTU - rok 2007

4.1.1. POPIS ŘEŠENÍ PROJEKTU - seznam

	Pořadí	Soubor
	1	**Postup řešení projektu v roce 2007 - Plzeň** Soubor obsahuje přehled nejvýznamnějších výsledků řešení projektu dosažených v průběhu roku 2007. Všechny vytýčené cíle byly splněny, pro jejich naplnění však bylo třeba provést celou řadu dodatečných činností, které jsou z důvodu rozsáhlosti zprávy v odstavci popsány pouze částečně. Detailní výsledky je možno nalézt v přílohách. Soubory-Zpravy-2C06009-PRI0411-Zprava_2C06009_odst411.doc *(153 kB)*
	2	**Postup řešení projektu v roce 2007 - Brno** Zprava_2C06009_odst411_Brno.doc *(47 kB)*

4.1.2. DOSAŽENÉ VÝSLEDKY

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/01/2007**

Název výsledku

Využití N-gramů pro odhalování plagiátů

Abstrakt

Stále rostoucí popularita Internetu a zvyšující se dostupnost různých dokumentů nám přináší i jisté problémy. Jedním z mnoha příkladů je množství pokusů o kopírování cizích prací, s vizí ulehčit si vlastní námahu. To s sebou přináší i rozvoj metod jak plagiátora identifikovat. Tento článek objasňuje metody pro detekci plagiátů a přibližuje náš výzkum v současnosti probíhající na ZČU. Zájem je věnován především metodě využívající n-gramy pro detekci překrývajících se částí dokumentů a odstranění problémů s posunem textu. K extrakci n-gramů vyšších řádů jsou na konci článku porovnány různé metody.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

V tomto článku jsme nastínili problematiku plagiátorství a využití n-gramů pro odhalování plagiátů. Zároveň byly objasněny různé varianty aplikace n-gramů na straně ověřovaného dokumentu a databáze, včetně jejich výhod i nevýhod. Pro extrakci vyšších řádů n-gramů a výpočet jejich četností jsme provedli srovnání metod Suffix Tree, Suffix Array a invertovaného indexu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Z výsledků je patrné, že metoda Suffix Array dosahuje téměř konstatního času pro různé řády, a proto je lepší volbou pro naše experimentování.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Češka Z.:Využití N-gramů pro odhalování plagiátů. In Proceedings of the ITAT 2007, Information Technologies - Applications and Theory, pp. 63-66, Polana, Slovakia, September 2007. ISBN 978-80-969184-6-1.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/02/2007**

Název výsledku

Teraman: A Tool for N-gram Extraction from Large Datasets

Abstrakt

In natural language processing (NLP) mainly single words are utilized to represent text documents. Recent studies have shown that this approach can be often improved by employing other, more sophisticated, fea-tures. Among them, mainly N-grams have been success-fully used for this purpose and many algorithms and procedures for their extraction have been proposed. However, usually they are not primarily intended for large data processing, which has currently become a critical task. In this paper we present an algorithm for N-gram extraction from huge datasets. The experiments indicate that our approach reaches outstanding results among other available solutions in terms of speed and amount of processed data.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Metoda umožňující extrakci N-gramů z rozsáhlých textových dat na jednom počítači bez použití specializovaného hardware. Námi navržená metoda překonává ostatní metody ve smyslu časových i paměťových požadavků a splňuje kritéria kladená na množství zpracovaných dat na webu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přínosy jsou především ekonomického charakteru, protože zde nabízíme metodu dovolující zpracování rozsáhlých dat na jednom běžném počítači bez dodatečných investic do hardwarového vybavení.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Teraman: A Tool for N-gram Extraction from Large Datasets	S - Prototyp, uplatněná metodika, funkční vzorek, autorizovaný software, výsledky aplikovaného výzkumu promítnuté do právních předpisů a norem, užitný vzor	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/03/2007**

Název výsledku

The Future of Copy Detection Techniques

Abstrakt

Internet is one of the richest encyclopaedias in the world. Students can easily download various free documents and then plagiarize their content. This paper describes the current state of copy detection methods and proposes some new trends. New approaches, closer to nature language processing, can essentially improve identification of hardly-detectable cases of plagiarism, i.e. single word changes and sentence structure changes. Synonyms and Latent Semantic Analysis are discussed in detail for better understanding of the semantics within documents.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití metody SVD pro detekci plagiátů prostřednictvím překrývajících se frazí, které jsou reprezentovány n-gramy.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Návrh nové metody detekující plagiáty na základě skrytých sémantických vlastností textu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Češka Z., Hanák I., Tesař R.: A Tool for N-gram Extraction from Large Datasets. In Proceedings of the IEEE 3rd International Conference on Intelligent Computer Communication and Processing (IEEE ICCP 2007), pp. 209-216, Cluj-Napoca, Romania, September 2007. ISBN 978-1-4244-1491-8.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/04/2007**

Název výsledku

Vliv normalizace slov na klasifikaci textů

Abstrakt

Práce porovnává vliv různých normalizačních metod na klasifikační úlohu. Část článku je věnována popisu naší lemmatizační metody založené na použití tezauru EWN. Prezentujeme srovnání výsledků získaných EWN metodou a ostatními normalizačními metodami. Zkoumána je také celková míra ovlivnění výsledků klasifikace textu jeho předzpracováním – normalizací slov a odstraněním stop-slov.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Metoda EWN se jeví jako vhodná pro lemmatizaci textů a je svými vlastnostmi unikátní. Obě konfigurace – s použitím a bez použití synsetů (množiny synonym) – produkují slibné výsledky na testovacím korpusu ČTK. Očekáváme, že rostoucí kvalita tezauru EWN bude vylepšovat také výsledky lemmatizační metody.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Algoritmus ověřuje přístupy k normalizaci slov. Bylo zjistěno, že jako nejlepší přístup pro zpracování textových dokumentů se jeví odstraňování stop-slov bez použití normalizačních metod. Pokles přesnosti klasifikace v případě aplikování normalizace je zřejmý a v některých případech statisticky významný. Na druhou stranu umožňuje normalizace redukci korpusu a dimenze dokumentů, což je přínosné, preferujeme-li zpracování velkého objemu dat nebo rychlost.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Toman Michal Ing.**
Spojení	+420377632452 mtoman@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Toman M.,Tesař R., Ježek K.: Vliv normalizace slov na klasifikaci textů. In Znalosti 2007, Ostrava: VŠB - Technická univerzita, Czech Republic, ISBN 978-80-248-1279-3, pages 360-363, February 2007.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/05/2007**

Název výsledku

Text Summarization within the LSA Framework

Abstrakt

This thesis deals with the development of a new text summarization method that uses the latent semantic analysis (lsa). The language-independent analysis is able to capture interrelationships among terms, so that we can obtain a representation of document topics. This feature is exploited by the proposed summarization approach. The method originally combines both lexical and anaphoric information. Moreover, anaphora resolution is employed in correcting false references in the summary. Then, I describe a new sentence compression algorithm that takes advantage from the lsa properties. Next, I created a method which evaluates the similarity of main topics of an original text and its summary, motivated by the ability of lsa to extract topics of a text. Using summaries in multilingual searching system muse led to better user orientation in the retrieved texts and to faster searching when summaries were indexed instead of full texts.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

- nová metoda sumarizace textů a její další vylepšení - nová metoda hodnocení kvality souhrnů

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

- použití nové metody sumarizace v multilingválním vyhledávacím systému – lepší orientace uživatele a rychlejší vyhledávání

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD**
Spojení	377632401 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J.: Text Summarization within the LSA Framework. Doctoral Thesis, Pilsen 2007	O - Ostatní výsledky, které nelze zařadit do žádného z výše uvedených druhů výsledku	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/06/2007**

Název výsledku

Two uses of anaphora resolution in summarization

Abstrakt

We propose a new method for using anaphoric information in Latent Semantic Analysis (LSA), and discuss its application to develop an LSA-based summarizer which achieves a significantly better performance than a system not using anaphoric information, and a better performance by the ROUGE measure than all but one of the single-document summarizers participating in DUC-2002. Anaphoric information is automatically extracted using a new release of our own anaphora resolution system, GUITAR, which incorporates proper noun resolution. Our summarizer also includes a new approach for automatically identifying the dimensionality reduction of a document on the basis of the desired summarization percentage. Anaphoric information is also used to check the coherence of the summary produced by our summarizer, by a reference checker module which identifies anaphoric resolution errors caused by sentence extraction.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

-nová metoda použití rezoluce anafor při sumarizaci textů -nová metoda kontroly anafor v souhrnu

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

-vylepšení extrakce vět (použití anafor) -vylepšení kvality souhrnu (kontrola anafor)

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. Phd.**
Spojení	377632401 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Poesio M., Kabadjov M. A., Jezek K.: Two Uses of Anaphora Resolution in Summarization. Information Processing & Management , Elsevier Ltd, Vol.43,Issue6, November 2007	J - Článek v odborném periodiku	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/07/2007**

Název výsledku

Identifying Novel Information Using Latent Semantic Analysis

Abstrakt

In our two-stage system for the English monolingual WiQA Task, snippets were first retrieved if they contained an exact match with the title. Candidates were then passed to the Latent Semantic Analysis component which judged them Novel if their match with the article text was less than a threshold. In Run1, the ten best snippets were returned and in Run 2 the twenty best. Run 1 was superior, with Average Yield per Topic 2.46 and Precision 0.37. Compared to other groups, our performance was in the middle of the range except for Precision where our system was the best. We attribute this to our use of exact title matches in the IR stage. In future work we will vary the approach used depending on the topic type, exploit co-references in conjunction with exact matches and make use of the elaborate hyperlink structure which is a unique and most interesting aspect of the Wikipedia.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

-nová metoda vyhledávání dalších informací k dokumentům a rozhodování, zda jsou tyto informace nové nebo redundantní

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

-vyhledání dalších informací o subjektech ve wikipedii

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377632401 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Sutcliffe R. F. E., Steinberger J., Kruschwitz U., Poesio M., Kabadjov M.A.: Identifying Novel Information using Latent Semantic Analysis. Evaluation of Multilingual and Multi-modal Information Retrieval. LNCS 4730, pp 541-549, Springer 2007, ISBN 3-540-74998-1	C - Kapitola v knize	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/08/2007**

Název výsledku

LSA-Based Multi-Document Summarization

Abstrakt

We present our first multilingual multi-document summarizer. The proposed LSA-based method satisfies the multilinguality constraint because it works only with the context of terms. We experiment with Czech and English Corpora. The experiments show that the summarizer is comparable with the best DUC participated systems.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Rozšíření metody sumarizace založené na LSA – sumarizace shluku dokumentů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Vícejazyčná podpora, sumarizace shluku dokumentů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. Ph.D.**
Spojení	377632401 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Křišťan M.: LSA-Based Multi-Document Summarization. In Proceedings of the 8th International Workshop on Systems and Control, a Young Generation Viewpoint, Balatonfured, Hungary, September 2007, pp. 97-101, ISBN 978-963-311-365-3.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/09/2007**

Název výsledku

Knowledge-poor Multilingual Sentence Compression

Abstrakt

We present a feature-based method for sentence compression. Firstly, a summary is created by our summarization method based on latent semantic analysis. The compression approach then removes unimportant clauses from the summary sentences. For each sentence a set of its possible compressed forms (compression candidates) is created. The candidates are then classified using 8 proposed features into two classes: in the first class there are candidates in which the important information was removed by compression and in the second class the information was still contained. The shortest candidate from the latter group substitutes the full sentence in the summary. The features are knowledge-poor which enables them to work with whatever language and the method can be easily extended by other features.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Nová metoda komprese vět pro sumarizátor textů založený na LSA.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Komprese vět v souhrnu (další zestručnění), vícejazyčná podpora.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. Ph.D.**
Spojení	377632401 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.textmining.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Steinberger J., Tesař R.: Knowledge-poor Multilingual Sentence Compression. In Proceedings of the 7th Conference on Language Engineering, Cairo, Egypt, December 2007, pp. 369-379.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/10/2007**

Název výsledku

The Java Abstract Annotation

Abstrakt

Recent trends in NLP (Natural Language Processing) are heading towards a stochastic processing of natural language. Stochastic methods, however, usually demand a lot of annotated training data. In most cases, the annotation of the data has to be done manually by a team of annotators and it is a highly time-consuming and expensive process. Thus we tried to develop an efficient and user-friendly editor that would aid human annotators to create the annotated data. We offer this editor for free. The developed editor is described in this article.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

User-friendly editor umožňující anotátorům vytvářet anotovaná data.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Zlepšení a urychlení práce při vytváření korpusů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Habernal Ivan Ing.**

Spojení 377 632 491 377 632 402 habernal@kiv.zcu.cz

Organizace 49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Habernal, I.; Konopík, M.: JAAE: The Java Abstract Annotation. In Proceedings of Interspeech 2007. Bonn: ISCA, 2007. s. 1298-1301. ISSN 1990-9772.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/11/2007**

Název výsledku

The Semantic Range of Spoken Dialogue Systems

Abstrakt

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Mouček Roman Ing. PhD.**
Spojení	377 632 441 377 632 402 moucek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Mouček, R.; Konopík, M.: The Semantic Range of Spoken Dialogue Systems. In: SPECOM 2007 proceedings. Moskva: Moscow State Linguistic University, 2007, s. 720-724. ISBN 6-7452-0110-X.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/12/2007**

Název výsledku

Towards Semantic Analysis of Spoken Queries

Abstrakt

In this article we describe an approach to semantic analysis of spoken queries to an internet information retrieval engine. In our system, a user will be enabled to express his or her requests by the natural language. Our system processes such a query and returns an appropriate answer from internet. In some domains, the approach that uses sentences rather than keywords can surpass systems that use just keywords to express the query. Our approach is based on the local parsing of lexical classes followed be a stochastic parsing algorithm.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

A special approach to semantic analysis of spoken queries to an internet information retrieval engine.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Obecně usnadnění vyhledávání položek pomocí internetových vyhledávačů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing.**
Spojení	377 632 491 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Konopík, M.; Mouček, R.: Towards Semantic Analysis of Spoken Queries. In: SPECOM 2007 proceedings. Moscow: Moscow State Linguistic University, 2007, s. 817-822. ISBN 6-7452-0110-X.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/13/2007**

Název výsledku

Voice controlled chess usable also for blind people

Abstrakt

This article deals with modification of a chess program so that it can be used by blind people trough voice control in natural Czech language. Open source software freely distributed under GNU General Public License (GPL) named jChecs is used as a chess engine and GUI. The chess game is controlled by a general purpose dialog manager. Voice recognizer named JLaser, the dialog manager and speech concatenative synthesis based on sentence phrases is added to the JChecs program.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Control of a chess engine by a general purpose dialog manager.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Modification of a chess program so that it can be used by blind people trough voice control in natural Czech language.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Hošna Martin Ing.**
Spojení	377 632 491 377 632 402 rekurze@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Hošna, M.; Konopík, M. Voice Controlled Chess usable also for Blind People. In: SPECOM 2007 Proceedings. Moskva: Moscow State Linquistic University, 2007, s. 725-728. ISBN 6-7452-0110-X.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/14/2007**

Název výsledku

Analysis of User Queries to the Internet

Abstrakt

The topic of this article is a part of the COT-SEWing project. The purpose of the COT-SEWing project is to develop a complex base of tools which will remove some of the typical barriers present in communication between human user and computerwithin the scope of Internet access. The necessity to express queries to an Internet search engine in keywords limits both the user and the searching system. We plan to remove this barrier by developing a system that would be able to analyze the meaning of spoken queries. Internet users nowadays are accustomed to use mostly keywords to express the query to an Internet search engine. In this article we describe an experiment that tries to find out whether users still sometimes use whole sentences to express queries. We also need a collection of whole sentence queries for our research. Thus we created a classifier that distinguishes between keyword queries and queries that have the form of a whole sentence. Such classifier would help us to filter user queries to search engine in order to find the queries that have the form of whole sentences. The goal of the COT-SEWing project is to deal with the Czech language we therefore concentrate on Internet queries in Czech.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Development of a complex base of tools which will remove some of the typical barriers present in communication between human user and computer within the scope of the Internet access.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Creation of an experiment that tries to find out whether users still sometimes use whole sentences to express queries.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing.**
Spojení	377 632 491 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Konopík, M.: Analysis of User Queries to the Internet. In: Proc. of PHDWS 2007, Computer and Automation Research Institute, Budapest, Hungary, 2007	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/15/2007**

Název výsledku

JLASER: An Automatic Speech Recognizer Written in Java

Abstrakt

We present an implementation of our automatic speech recognizer in Java programming language. The system supports neural network based acoustic models as well as the more traditional continuous density hidden Markov models and can read HTK trained models. Back-off n-gram language models and regular grammars may be used during recognition. The sources of the recognizer will be made available from our web pages liks.fav.zcu.cz.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

A novel special implementation of a very accurate automatic speech recognizer.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Velmi spolehlivé automatické rozpoznávání jednoduchých i mírně složitých promluv.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 377 632 402 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Pavelka, T., Ekštein, K.: JLASER: An Automatic Speech Recognizer Written in Java, Proc. of XII International Conference Speech and Computer (SPECOM'2007), Moscow, Russia, 2007, ISBN 6-7452-0110-X.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/16/2007**

Název výsledku

Context Dependency in Neural Network Based Acoustic Models

Abstrakt

Our recent experiments with Gaussian mixture (GMM) based acoustic models have shown that employing context dependent acoustic models, namely triphones, can greatly improve recognition accuracy in comparison to systems based on context independent units. Significant portion of our research has been aimed at exploring the possibilities of neural networks as acoustic models for speech recognition. We have observed, that a neural network can lead to similar recognition accuracy as a GMM acoustic model, while having less trainable parameters. This article discusses the use of decision tree clustered triphones represented by output neurons of a multi layer perceptron.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Special method of a speech recognition by a multilayer perceptron - the attempt how to use the neural networks for the speech recognition.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Development of a special kind of a speech recognizer and its use for reliable speech recognition.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 377 632 402 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Pavelka, T., Hejtmánek, J.: Context Dependency in Neural Network Based Acoustic Models, Proc. of PhD Workshop 2007, Budapest, Hungary, 2007	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/17/2007**

Název výsledku

Use of context-dependent units in Czech speech

Abstrakt

This work extends the LASER (ASR system which is being developed at the Laboratory of Intelligent Communication Systems ZČU, CZ) with the context-dependent units. It also explores the problems and their solving that comes with the implementation of context-dependent units into the recognizer. We explored two main methods of parameter clustering: Data-driven and Decision tree based. The clustering methods brought very good results and will be in the spot of our future research. After all tests we assume that using context-dependent units in computer speech recognition it is possible to gain even better results.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Two main methods of parameter clustering: data-driven and decision tree based.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Development of a new kind of a very reliable speech recognizer.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Hejtmánek Jan Ing.**
Spojení	377 632 491 377 632 402 hejtman2@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Hejtmánek, J., Pavelka, T.: Use of context-dependent units in Czech speech. Proc. of PhD. Workshop 2007, Balatonfüred, Maďarsko, 2007	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/18/2007**

Název výsledku

Confidence Measures for Semi-automatic Labeling of Dialog Acts

Abstrakt

This paper deals with semi-supervised classifier training for automatic Dialog Acts (DAs) recognition. In our previous works, we have designed a dialog act recognition system for reservation applications in the Czech language. In this work, we propose to retrain this system on another corpus, for another task (broadcast news speech), in a different language (French) and with another set of dialog acts. This is realized using a semi-supervised approach based on the Expectation-Maximization (EM) algorithm. We show that, in the proposed experimental setup, the use of confidence measures to filter out incorrectly recognized dialog acts is required to improve the results. Two confidence measures are thus proposed and evaluated on the French broadcast news corpus. Experimental results confirm the interest of this approach for the task of training automatic dialog act classifiers.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Development of a semi-supervised classifier training method for semi-automatic Dialog Acts (DAs) corpora creation.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

We proposed to retrain the recognition system on another corpus, for another task (broadcast news speech), in a different language (French) and with another set of dialog acts. It was realized using a semi-supervised approach based on the Expectation-Maximization (EM) algorithm.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing. PhD.**
Spojení	377 632 495 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kral, P., Cerisasa, C., Kleckova, J.: Confidence Measures for Semi-automatic Labeling of Dialog Acts. In: Proc. of ICASSP'07, Honolulu, Hawaii, USA, April 2007, pp. 153-156, ISSN 1520-6149	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/19/2007**

Název výsledku

Lexical Structure for Dialogue Act Recognition

Abstrakt

This paper deals with automatic dialogue acts (DAs) recognition in Czech. Dialogue acts are sentence-level labels that represent different states of a dialogue, such as questions, hesitations, ... In our application, a multimodal reservation system, four dialogue acts are considered: statements, orders, yes/no questions and other questions. The main contribution of this work is to propose and compare several approaches that recognize dialogue acts based on three types of information: lexical information, prosody and word positions. These approaches are tested on a Czech Railways corpus that contains human-human dialogues, which are transcribed both manually and with an automatic speech recognizer for comparison. The experimental results confirm that every type of feature (lexical, prosodic and word positions) bring relevant and somewhat complementary information. The proposed methods that take into account word positions are especially interesting, as they bring global information about the structure of the sentence, at the opposite of traditional n-gram models that only capture local cues. When word sequences are estimated from a speech recognizer, the resulting decrease of accuracy of all proposed approaches is very small (about 3%), which confirms the capability of the proposed approaches to perform well in real applications.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

To propose and compare several approaches that recognize dialogue acts based on three types of information: lexical information, prosody and word positions.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Bringing global information about the structure of the sentence, at the opposite of traditional n-gram models that only capture local cues.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing.**
Spojení	377 632 496 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kral, P., Cerisasa, C., Kleckova, J.: Lexical Structure for Dialogue Act Recognition. In: Journal of Multimedia (JMM), Volume 2, Issue 3, June 2007, pp. 1-8, ISSN 1796-2048	J - Článek v odborném periodiku	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/20/2007**

Název výsledku

Importance of Prosody for Dialogue Act Recognition

Abstrakt

This paper deals with automatic Dialogue Acts (DAs) recognition in French and in Czech. In this work, only prosodic features are considered. The utterances are recognized according to three types of dialogue acts: statements, yes/no questions and other questions, mainly wh-questions. We show that it is not possible to recognize all utterances only with basic prosodic features (F0 and energy) in real conditions with a good accuracy.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

We show in this article that it is not possible to recognize all utterances only with basic prosodic features (F0 and energy) in real conditions with a good accuracy.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

An improvement of DA recognition rate

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing.**
Spojení	377 632 496 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kral, P., Cerisara, C., Kleckova, J.: Importance of Prosody for Dialogue Act Recognition. In: Proceedings of SPECOM'07, Moscow, Russia, October 2007, pp. 140-145, ISBN 6-7452-0110-X	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/21/2007**

Název výsledku

Automatic Recognition of Dialogue Acts

Abstrakt

This thesis deals with automatic Dialogue Act (DA) recognition in Czech and in French. Dialogue acts are sentence-level labels that represent different states of a dialogue, such as questions, statements, hesitations, etc. The first main contribution of this work is to propose and compare several approaches that recognize dialogue acts based on three types of information: lexical, prosodic and word positions. These approaches are tested on the Czech Railways corpus that contains human-human dialogues, which are transcribed both manually and with an automatic speech recognizer for comparison. The experimental results confirmed that every type of feature (lexical, prosodic and word positions) bring relevant and somewhat complementary information. The proposed methods that take into account word positions are especially interesting, as they bring global information about the structure of a sentence, at the opposite of traditional n-gram models that only capture local cues. One of the main issue in the domain of automatic dialogue act recognition concerns the design of a fast and cheap method to label new corpora. The next main contribution is to apply the general semi-supervised training approach based on the Expectation Maximization algorithm to the task of labeling a new corpus with the pre-defined DAs. We further proposed to filter out the examples that might be incorrect by two confidence measures, namely the maximum a posteriori probability and the a posteriori probability difference methods. Experimental results showed that the proposed method is an efficient approach to create new dialogue act corpora at low costs.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

The main result of this work is the proposal and comparison of several approaches that recognize dialogue acts based on three types of information: lexical, prosodic and word positions. The experimental results confirmed that every type of feature (lexical, prosodic and word positions) bring relevant and somewhat complementary information.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The main contribution of this work is to take word positions into account, as they bring global information about the structure of a sentence, at the opposite of traditional n-gram models that only capture local cues. The next main contribution of this work is to apply the general semi-supervised training approach based on the Expectation Maximization algorithm to the task of labeling a new corpus with the pre-defined DAs.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing.**
Spojení	377 632 496 377 632 402 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Kral, P.: Automatic Recognition of Dialogue Acts. CoPhD. Thesis, Henri Poincaré University – Nancy 1 and University of West Bohemia in Pilsen, November 2007	O - Ostatní výsledky, které nelze zařadit do žádného z výše uvedených druhů	ANG

výsledku

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/22/2007**

Název výsledku

The Java Abstract Annotation Editor

Abstrakt

Česky: V současnosti se v oblasti zpracování řeči začínají stále více používat stochastické metody. Ty však většinou vyžadují velké množství anotovaných trénovacích dat. Popsaný editor je navržen tak, aby maximálně pomohl anotátorům při značkování dat. Anglicky: Recent trends in NLP are heading towards a stochastic processing of natural language. Stochastic methods usually demand a lot of annotated data. Thus we tried to develop an efficient editor that would aid human annotators to create the annotated data.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Sémantické anotace bývají obvykle vytvářeny anotátory ručně v textovém editoru. Navržená metodologie a softwarový produkt pro podporu tvorby sémantických anotací značně usnadňuje a urychluje proces vytváření sémantických anotací. Vytvořený editor sémantických anotací využívá GUI (grafické uživatelské rozhraní) a definici anotačních schémat (XML datový soubor) a snižuje tak náklady na zdroje potřebné pro získání sémantických anotací. Inovačním aspektem je nový přístup k pořizování sémantických anotací.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Editor do značné míry usnadňuje a zrychluje vytváření trénovacích dat pro algoritmy sémantické analýzy. Takto lze snížit obyčejně vysoké náklady nutné pro pořízení těchto dat.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing. .**
Spojení	377 632 491 377 632 402 kekstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Habernal, I.; Konopík, M. JAAE: The Java Abstract Annotation Editor. In: Proceedings of Interspeech 2007. Bonn: ISCA, 2007, s. 1298-1301, ISSN 1990-9772	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG
02	Dokumentace je dostupná na: http://liks.fav.zcu.cz/mediawiki/index.php/JAAE	A2 - Prezentace v oblasti VaV - elektronický dokument se vzdáleným přístupem	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/23/2007**

Název výsledku

JLASER - Java LICS Automatic Speech Extraction/Recognition

Abstrakt

Implementace systému automatického rozpoznávání řeči JLASER v jazyce Java přináší řadu výhod vyplývajících z objektově orientovaného návrhu, vyšší úrovně abstrakce jazyka a velikosti komunity vývojářů v programujících v Javě. Nevýhodou je rychlost běhu, která může být až o polovinu menší než při implementaci v jazyce nižší úrovně, jako je např. C. Při rozpoznávání je možné použít modely natrénované nástrojem HTK (Hidden Markov Toolkit). Mimo vlastního rozpoznávače systém poskytuje nástroje pro automatické značkování trénovacích nahrávek, automatickou ortograficko-fonetickou transkripci (pro češtinu), tvorbu rozpoznávacích grafů ze slovníků a gramatik a měření úspěšnosti rozpoznávání.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- JC, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Kromě tradičního přístupu k akustickému modelování založeného na směsi Gaussových mixtur je podporován i tzv. hybridní přístup, který kombinuje výhody skrytých Markovových modelů a umělých neuronových sítí.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Vzhledem k tomu, že software je napsaný v jazyce Java, je nezávislý na platformě, snadno upravitelný, rozšiřitelný a integrovatelný do programů napsaných v Javě.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 377 632 402 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Dokumentace je dostupná na: http://liks.fav.zcu.cz/mediawiki/index.php/JLASER	A2 - Prezentace v oblasti VaV - elektronický dokument se vzdáleným přístupem	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/24/2007**

Název výsledku

LAC-SS (LICS Audio Corpus/Spontaneous Speech)

Abstrakt

Korpus LAC-SS je korpus nahrávek spontánní řeči opatřený rozšířenou synchronní ortografickou transkripcí. V současné době obsahuje LAC-SS celkem 741 minut (12h21m45s) záznamu spontánní řeči ve 24 nahrávkách vysoké kvality, k nimž byla pořízena manuální transkripce. 11 nahrávek pochází z přednášek pracovníků Katedry informatiky a výpočetní techniky - jedná se tedy o mluvčí s rozsáhlou praxí v oblasti veřejných projevů. Naopak 13 nahrávek pochází ze studentských seminářů - mluvčí (studenti) ve většině případů nejsou zvyklí pronášet veřejné projevy, a proto také záznamy obsahují celou řadu neřečových zvuků a projevů, např. nervozity, což je v tomto případě ale žádoucí, protože díky tomu je korpus dostatečně foneticky bohatý. Z pohledu výpočetní lingvistiky korpus obsahuje celkem 40866 lexikálních atomů, z toho 33795 slov a 7071 neřečových zvuků (non-speech sounds, např. odkašlávání, popotahování, kýchání, mlaskání, apod.). Celkový počet různých slov v korpusu je 6894.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Značné množství současných konkurenčních ASR systémů je trénováno daty připravenými uměle, tj. nahrávkami profesionálních rečníků v ideálních příjmových podmínkách, v zatlumených studiích, kvalitními mikrofony, atd. Korpus LAC-SS umožňuje natrénovat rozpoznávač reálnými daty, přičemž nejvýznamnější přínos není ani tak přítomnost projevů akustického prostředí, ale zejména hojná existence artefaktů spontánní řeči: přeřeknutí, zakoktání, falešných začátků, atd.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Značné zvýšení spolehlivosti rozpoznávání v reálných podmínkách či idealizovaných reálných podmínkách. Přítomnost neřečových zvuků a artefaktů spontánní řeči také umožňuje trénovat tzv. garbage modely rozpoznávače, díky nimž je ASR systém odolných vůči takovým projevům a dokáže je správně zpracovat (tj. ignorovat). Dalším významným přínosem je vytvoření vysoce přesné synchronní ortografické transkripce korpusového materiálu, díky které je možné snadno a rychle pomocí výpočetních techniky provádět analýzy a statistiky spontánní řeči.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ekštein Kamil Ing. PhD.**
Spojení	377 632 491 377 632 402 kekstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Podrobnější statistika získaného materiálu je k dispozici v článku: Ekštein, K.: On Building of Czech Spontaneous Speech Corpus. Proc. of XII International Conference Speech and Computer (SPECOM'2007), Moscow, Russia, 2007, ISBN 6-7452-0110-X.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/25/2007**

Název výsledku

Rozšíření bag-of-words modelu dokumentu: srovnání bigramů a 2-itemsetů.

Abstrakt

Při kategorizaci textu lze reprezentovat dokumenty jednotlivými slovy. Tento přístup je označován jako bag-of-words nebo také single words-based. Nicméně dalším obohacením této reprezentace je možné dosáhnout zlepšení výsledků klasifikace. V této práci jsme zaměřili svou pozornost na porovnání přínosu bigramů a 2-itemsetů, o které je rozšířen klasický bag-of-words model dokumentu. K experimentům využíváme standardní anglické textové korpusy Reuters-21578 a 20 Newsgroups. Ke klasifikaci je použit multinomial Naive Bayes, protože pro tuto klasifikační metodu a výše zmíněné korpusy byla publikována řada odborných publikací, se kterými naše dosažené výsledky srovnáváme. K výběru charakteristických položek (feature selection) využíváme 5 různých přístupů. Naše experimenty indikují, že použitím bigramů a 2-itemsetů je možné statisticky významně zvýšit úspěšnost klasifikace. Dále je v případě 2-itemsetů velmi důležité zvolit vhodný způsob výběru charakteristických položek. Na druhou stranu, v případě bigramů je možné dosáhnout zlepšení úspěšnosti klasifikace i použitím velmi jednoduchého přístupu. Z našich experimentů usuzujeme, že není příliš efektivní rozšiřovat reprezentaci textového dokumentu o 2-itemsety, protože pomocí bigramů je možné dosáhnout lepších výsledků a jejich generování je oproti 2-itemsetům méně náročné.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Obohacení klasického modelu textového dokumentu založeného na jednotlivých slovech o bigramy nebo 2-itemsety umožňuje zlepšit úspěšnost klasifikace, v některých případech dokonce statisticky významně. Prokázání, že bigramy jsou oproti 2-itemsetům pro obohacení modelu dokumentu vhodnější, průměrně dosahují většího zlepšení při současné nižší složitosti jejich generování z textu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přínos pro uživatele představuje prokázání skutečnosti, že obohacením modelu dokumentu o bigramy i 2-itemsety je možné dosáhnout statisticky významného zlepšení výsledků klasifikace oproti běžnému modelu založenému jen na samostatných slovech. Důležitý přínos představuje znalost, že bigramy jsou pro tento účel oproti 2-itemsetům vhodnější.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Tesař Roman Ing**

Spojení 377632479 romant@kiv.zcu.cz

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
00	Tesař R., Poesio M., Strnad V., Ježek K.: Rozšíření bag-of-words modelu dokumentu: srovnání bigramů a 2-itemsetů. In: Sborník konference Znalosti 2007, pp.131-142, ISBN 978-80-248-1279-3, Ostrava.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/26/2007**

Název výsledku

The Fight against Spam - A Machine Learning Approach

Abstrakt

The paper presents a survey of the duel between spammers and antispam software developers, and also describes new approaches to spam filtering. In the first two sections we present a survey of the currently existing spam types. Some well-mapped spammer tricks are also described, although the imagination of spam distributors is endless, and therefore only the most common tricks are covered. We present some up-to-date spam blocking techniques currently integrated into today's spam filters. In Methodology and Results sections we describe our implementation of Itemsets-based, Naïve Bayes and LSI classifiers for classifying email messages into spam and non-spam (ham) categories.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Článek nejen poskytuje informace o současném stavu a trendech v oblasti spamových a antispamových technik, popisuje i návrh, implementaci a otestování vlastních antispamových filtrů, založených na Itemsetech, na NB klasifikátoru a na LSI klasifikátoru.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Výsledky jsou využitelné v případě konstrukce spamových filtrů. Byla ověřena užitečnost a snadná aplikovatelnost Itemsetů a MB metody.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Ježek Karel Doc. Ing. CSc.**

Spojení 377632475 jezek_ka@kiv.zcu.cz

Organizace 49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
00	Jezek K., Hynek J.: The Fight against Spam - A Machine Learning Approach. In Proceedings of the 11th International Conference on Electronic Publishing, pp.381-392, ISBN 978-3-85437-292-9, Vienna, Austria	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/27/2007**

Název výsledku

SPOT – nový webový projekt on-line slovníku překladů odborných termínů

Abstrakt

Slovník odborné terminologie slouží dvěma účelům: jako zdroj referenčních překladů a jako platforma pro diskuse při jejich „ustalování“. SPOT bude rovněž možné nasadit v dalších aplikačních oblastech, jako je sémantický web, multilinguální rozhraní, rozšiřování dotazů v rámci IR, thesaury apod.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Zobrazování originálních i překladových termínů / kolokací v různém kontextu (www, domény www, online slovníky, blogy apod.) Online diskuse a hlasování o různých formách překladu odborných výrazů

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Zkrácení času při posuzování vhodnosti konkrétní volby překladu odborného výrazu Možnost podílet se na ustalování odborné terminologie Možnost sledovat historii vývoje termínu/kolokace Provádění dávkových aktualizací korpusu Sdílení podmnožiny korpusu v rámci projektového týmu

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Hynek Jiří Ing. Phd.**
Spojení	377632401 603492837 jhynek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
00	Hynek J., Brada P.: On the Evolution of Computer Terminology and the SPOT On-Line Dictionary Project. In Proceedings of the 11th ICCC International Conference on Electronic Publishing, pp. 257-268, ISBN 978-3-85437-292-9, Vienna, Austria	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/28/2007**

Název výsledku

Ranking Algorithms For Web Sites: Finding Authoritative Academic Web Sites and Researchers

Abstrakt

In this paper, we discuss several common ranking algorithms for Web pages and we present a methodology based on them for finding authoritative researchers by analyzing academic Web sites. We show a case study in which we concentrate on a set of French computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones. We then examine the contents of the research papers present on these sites and determine the most authoritative French authors. We also propose some future improvements.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popisuje metodu pro hodnocení autoritativnosti institucí a osob z akademického prostředí na bázi referencí mezi web stránkami a její ověření na případu informatických fakult a kateder.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Metoda je využitelná při sestavování žebříčku kvality výukových a výzkumných subjektů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Fiala Dalibor Ing. PhD.**
Spojení	377632401 dalfia@kiv.zcu.cz; dalibor.fiala@gefasoft.de
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
00	Fiala D., Rousselot F., Ježek K. Ranking Algorithms For Web Sites: Finding Authoritative Academic Web Sites and Researchers.In Proceedings of the 3rd International Conference on Web Information Systems and Technologies WEBIST'07, pp. 372-375, ISBN 978-972-8865-78-8, Barcelona Spain 2007.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/29/2007**

Název výsledku

Využití struktury webu pro vyhledávání autoritativních institucí a osob

Abstrakt

V tomto článku představíme metodologii pro vyhledávání autoritativních vědeckých pracovníků analýzou webových stránek akademických pracovišť. Uvedeme případovou studii zaměřenou na skupinu stránek českých kateder informatiky. Nejprve zanalyzujeme odkazy mezi nimi (jejich vzájemné vztahy) a několika známými hodnoticími algoritmy stanovíme nejvýznamnější katedry. Potom prozkoumáme obsah výzkumných článků nalézajících se na těchto stránkách a určíme nejdůležitější české autory.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JDj, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Uvádí se případová studie aplikující metodu hodnocení akademických pracovišť a pracovníků na prostředí českých a francouzských informatických pracovišť. Rozšířená verze zdokonalená o hledisko koautorství a temporální aspekty vyjde v Scientometrics 1/2008

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Metodu lze aplikovat na libovolné akademické subjekty a získat tím jejich objektivní hodnocení.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Fiala Dalibor Ing. PhD.**
Spojení	377632401 dalfia@kiv.zcu.cz; dalibor.fiala@gefasoft.de
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
00	Fiala D., Jezek K., Rousselot F.:Využití struktury webu pro vyhledávání autoritativních institucí a osob. In Proc. 6th Annual Conf. ZNALOSTI 2007, pp. 300-303, ISBN 978-80-248-1279-3, Ostrava 2007	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/30/2007**

Název výsledku

Web Mining Methods for the Detection of Authoritative Sources

Abstrakt

The innovative portion of this doctoral thesis deals with the definitions, explanations and testing of modifications of the standard PageRank formula adapted for bibliographic networks. The new versions of PageRank take into account not only the citation but also the co-authorship graph. We verify the viability of the new algorithms by applying them to the data from the DBLP digital library and by comparing the resulting ranks of the winners of the ACM SIGMOD E. F. Codd Innovations Award. The rankings based on both the citation and co-authorship information turn out to be better than the standard PageRank ranking. In another part of the disseration, we present a methodology and two case studies for finding authoritative researchers by analyzing academic Web sites. In the first case study, we concentrate on a set of Czech computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones using several common ranking algorithms. We then examine the contents of the research papers present on these sites and determine the most authoritative Czech authors. In the second case study, we do exactly the same with French academic computer science Web sites to find the most significant French researchers in the field. We also discuss the weak points of our approach and propose some future improvements. To the best of our knowledge, it is the only attempt ever made at discovering authoritative researchers from the above countries by directly mining from Web data.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Adaptace PageRank formule pro její použití v bibliografických sítích při respektování různých vlivů koautorství na významnost citací.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Dovoluje objektivnější posuzování bibliografických a hyperlinkových citací.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Fiala Dalibor Ing. PhD.**
Spojení	377632401 dalfia@kiv.zcu.cz; dalibor.fiala@gefasoft.de
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Fiala D.: Web Mining Methods for the Detection of Authoritative Sources, PhD. Thesis, WBU Pilsen, LPU Strasbourg 2007	O - Ostatní výsledky, které nelze zařadit do žádného z výše uvedených druhů výsledku	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/31/2007**

Název výsledku

Konference Text, Speech and Dialogue 2007

Abstrakt

V září 2007 byl v Plzni uspořádán již desátý ročník výše citované konference, který se uskutečnil v nově vybudovaném konferenčním centru Primavera Plzeň. Konference se zúčastnilo celkem 136 účastníků z 38 zemí s 86 příspěvky. Sborník konference byl vydán nakladatelstvím Springer Verlag Berlin, Heidelberg pod signaturou LNAI 4629. Konference měla kromě odborného také společenský program, jenž měl dva zásadní cíle - umožnit úzký kontakt nových, mladých řešitelů výše zmíněné problematiky se špičkovými světovými představiteli oboru a dále pak představit světové veřejnosti oblast západních Čech a její rozvoj v posledním období.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Prezentace nejnovějších poznatků z oblasti zpracování textů a mluveného slova, vzájemná výměna zkušeností mezi řešiteli z celého světa a prezentace nejnovějšího programového vybavení jak pro zpracování textové informace, tak i pro analýzu a porozumění mluvenému slovu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Navázání úzkých kontaktů mezi řešiteli výše zmíněných úloh, seznámení se s nejnovějšími teoretickými i praktickými poznatky z oblastí počítačového zpracování textové informace a mluveného slova.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Matoušek Václav Prof. Ing. CSc.**
Spojení	377 632 471 377 632 402 matousek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz/tsd2007

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Konference Text, Speech and Dialogue, 2007	M - Uspořádání (zorganizování) konference	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/32/2007**

Název výsledku

Sborník konference TSD 2007

Abstrakt

The International Conference TSD 2007 presented state-of-the-art technology and recent achievements in the field of natural language processing. It declared its intent to be an interdisciplinary forum, intertwining research in speech and language processing with its applications in everyday practice. We feel that the mixture of different approaches and applications offered a great opportunity to get acquainted with the current activities in all aspects of language communication and to witness the amazing vitality of researchers from developing countries too.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Kompaktní knižní publikace obsahující veškeré příspěvky přednesené v rámci konference pořádané v září 2007 v Plzni.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Knižní publikace dostupná v celém světě, která přináší nejnovější poznatky z oblasti zpracování textové informace a počítačové analýzy a porozumění mluvenému slovu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Matoušek Václav Prof. Ing. CSc.**
Spojení	377 632 471 377 632 402 matousek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz/tsd2007

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Matoušek, V., Mautner, P.: Text, Speech and Dialogue - 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 2007. LNAI 4629, Springer Verlag, Berlin, Heidelberg, 2007, ISBN 978-3-540-74627-0.	B - Odborná monografie	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/33/2007**

Název výsledku

Sémantický korpus dotazů do internetového vyhledávače

Abstrakt

Sémantický korpus dotazů pořízený v průběhu roku 2007 obsahuje zatím celkem 20292 vět, z toho všechny věty jsou anotovány tématicky a dosud 6750 vět anotováno sémanticky.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JC, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Korpus pro analýzu promluv v přirozeném jayzce.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Korpus lze využít pro analýzu výstupů ze slovního recognizeru, popř. pro vytváření sémantických map.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing.**
Spojení	377 632 491 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Citace korpusu v: Konopík, M., Mouček, R.: Towards Semantic Analysis of Spoken Queries. In: SPECOM 2007 proceedings. Moscow: Moscow State Linguistic University, 2007. s. 817-822. ISBN 6-7452-0110-X.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/34/2007**

Název výsledku

Korpus reálných dotazů do internetového vyhledávače ve formě celých vět.

Abstrakt

V rámci aktivity byl vytvořen korpus reálných dotazů do internetového vyhledávače ve formě celých vět. Bylo vytvořeno 101 165 vět automatickou filtrací ze 187 miliónů dotazů získaných od firmy Seznam.cz. Všechny dotazy byly zpracovány týmem vyškolených anotačních pracovníků. Celá množina dotazů byla tématicky roztříděna a neužitečné dotazy byly odfiltrovány. Mezi anotatátorská shoda pro určování témat dosahuje 83%.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JC, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Byl získán nový korpus sémantických anotací pro český jazyk.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Korpus umožní vývoj algoritmů sémantické analýzy pro český jazyk.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing.**
Spojení	377 632 491 377 632 402 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Korpus byl představen v: Konopík, M.: Analysis of User Queries to the Internet. In: Proc. of PHDWS 2007, Computer and Automation Research Institute, Budapest, Hungary, 2007.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/35/2007**

Název výsledku

JSAPI: Java Speech API & Chart Parser Impelementation

Abstrakt

Recent trends in NLP (Natural Language Processing) are heading towards a stochastic processing of natural language. Stochastic methods, however, usually demand a lot of annotated training data. In most cases, the annotation of the data has to be done manually by a team of annotators and it is a highly time-consuming and expensive process.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JC, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Část systému pro automatické rozpoznávání řeči.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Tvorba nového typu řečového recognizeru, bude využit při inovaci návrhu systému.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing.**
Spojení	377 632 491 377 632 402 habernal@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Habernal I.: Sémantická analýza lexikálních tříd. Diplomová práce, KIV FAV ZČU v Plzni, 2007	O - Ostatní výsledky, které nelze zařadit do žádného z výše uvedených druhů výsledku	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/36/2007**

Název výsledku

Programový systém CzechWebSOM pro kategorizaci česky psaných dokumentů

Abstrakt

Implementace systému WEBSOM - kolekce metod pro zpracování sémantiky a kategorizaci kolekce českých dokumentů v jazyce Java přináší řadu výhod vyplývajících z objektově orientovaného návrhu, vyšší úrovně abstrakce jazyka a možnosti integrace softwarů třetích stran. Software umožňuje vytvoření číselné reprezentace kolekce dokumentů, redukci dimenze dat a zpracování dat dvouvrstvou architekturou Kohonenovy mapy (mapa slovních kategorií, mapa dokumentů). K dispozici je i grafické uživatelské rozhraní, které umožňuje uživatelsky jednoduché a příjemné nastavení parametrů metody a provedení většího množství experimentů.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Software je vyvíjen pro zpracování kolekcí článků v českém jazyce, zpracování dat je tedy přizpůsobeno českému jazykovému prostředí (např.lemmatizace). Použité algoritmy pro kategorizaci dokumentů je možné jednoduše parametrizovat vzhledem k rozsahu a povaze vstupních dokumentů, tudíž je možné nad kolekcí dokumentů provést velké množství experimentů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Software je napsaný v jazyce Java, je nezávislý na platformě, snadno upravitelný, rozšiřitelný a integrovatelný do dalších programů napsaných v tomto jazyce.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Mouček Roman Ing. PhD.**
Spojení	377 632 465 377 632 402 moucek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz/mediawiki/index.php/WEBSOM

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Dokumentace je dostupná na: http://liks.fav.zcu.cz/mediawiki/index.php/WEBSOM	S - Prototyp, uplatněná metodika, funkční vzorek, autorizovaný software, výsledky aplikovaného výzkumu promítnuté do právních předpisů a norem, užitný vzor	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/37/2007**

Název výsledku

Administration Framework for the DEB Dictionary Server

Abstrakt

This paper presents a new implementation of administration framework for the DEBII dictionary writing system. We present the details and examples of the user management part as well as graphical scenarios for dictionary service setup, adaptation and automatic generation of user application based on the dictionary XML schema.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš**
Spojení	+420-549 491 810 +420-549 491 820 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
37	Horák, Aleš - Rambousek, Adam. Administration Framework for the DEB Dictionary Server. In Computer Treatment of Slavic and East European Languages. Bratislava, Slovakia : Slovenská akadémia vied, 2007. od s. 70-79, 10 s. ISBN 9788087139059.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/38/2007**

Název výsledku

Building a Large Lexicon of Complex Valency Frames

Abstrakt

This paper describes the process of building and using a new comprehensive lexicon of Czech verb valency frames based on complex valency frames. The main features of the lexicon entries are designed to bring important semantic information to computer processing of predicate constructions in running texts. The most notable features include two-level semantic labels with linkage to the Princeton and EuroWordNet hierarchy and surface verb frame patterns used for automatic syntactic analysis. Some implications for other languages, particularly English, Bulgarian and Romanian, are reported.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel**
Spojení	+420-549 491 810 +420-549 491 820 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
38	Horák, Aleš - Pala, Karel. Building a Large Lexicon of Complex Valency Frames. In Proceedings of the FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages. Tartu, Estonia : Lund University, Sweden, 2007. od s. 31-38, 8 s. ISBN 978-91-976939-0-5.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/39/2007**

Název výsledku

DEB Platform tools for effective development of WordNets in application to PolNet.

Abstrakt

The objective of this paper is the presentation of DEB Platform Tools and their utility in creation of WordNet ontology resources. These tools have been elaborated at Masaryk University as a result of experiences coming from creation of Czech WordNet. This paper presents the benefits of introducing the tools into the process of production of Polish WordNet within the PolNet project (at Adam Mickiewicz University).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel**
Spojení	+420-549 491 810 +420-549 491 820 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
39	Pala, Karel - Horák, Aleš - Rambousek, Adam - Vetulani, Zygmunt - Konieczka, Paweł - Marciniak, Jacek - Obrębski, Tomasz - Rzepecki, Przemysław - Walkowska, Justyna. DEB Platform tools for effective development of WordNets in application to PolNet. In Proceedings of 3rd Language & Technology Conference. Poznań : Fundacja Uniwersytetu im. A. Mickiewicza, 2007. od s. 514-518, 5 s. ISBN 978-83-7177-407-2	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/40/2007**

Název výsledku

Derivational Relations in Czech WordNet

Abstrakt

In the paper we describe enriching Czech WordNet with the derivational relations that in highly inflectional languages like Czech form typical derivational nests (or subnets). Derivational relations are mostly of semantic nature and their regularity in Czech allows us to add them to the WordNet almost automatically. For this purpose we have used the derivational version of morphological analyzer Ajka that is able to handle the basic and most productive derivational relations in Czech. Using a special derivational interface developed in our NLP Lab we have explored the semantic nature of the selected noun derivational suffixes and established a set of the semantically labeled derivational relations presently 14. We have added them to the Czech WordNet and in this way enriched it with approx. 30 000 new Czech synsets.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel**
Spojení	+420-549 491 810 +420-549 491 820 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
40	Pala, Karel - Hlaváčková, Dana. Derivational Relations in Czech WordNet. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007. 1. vyd. Praha : The Association for Computational Linguistics, 2007. od s. 75-81, 6 s.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/41/2007**

Název výsledku

Dictionary Management System for DEB Development Platform.

Abstrakt

In the paper, we introduce new dictionary management interface for design, preparation and presentation of generic electronic XML dictionaries using the DEB (Dictionary Editing and Browsing) development platform. The DEB platform provides a strict client-server environment for general dictionary writing systems. So far several successful NLP tools have been implemented on this platform, one of the most known being the DEBVisDic tool for wordnet semantic network editing and visualization. This paper describes a new part of the DEB platform -- the Administration interface that is shared by all DEB applications running on one server machine.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš**
Spojení	+420-549 491 810 +420-549 491 820 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
41	Horák, Aleš - Rambousek, Adam. Dictionary Management System for DEB Development Platform. In NLPCS 2007: Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science. 2007. vyd. Funchal, Portugal : INSTICC PRESS, 2007. od s. 129-138, 10 s. ISBN 978-972-8865-97-9.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/42/2007**

Název výsledku

Displaying Bidirectional Text Concordances in KWIC format

Abstrakt

In the paper, we describe the problem of displaying bidirectional texts in the word concordance view and introduce a system that can handle these texts. A few examples of English word sequences in a corpus of Persian are given. We describe display algorithms and corpus input file modifications needed to achieve the correct word order in the concordance view. We also discuss some related problems, e.g. working with neutral characters (like punctuation or numbers) and the recognition of the left-to-right (right-to-left) text boundaries.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Rychlý Pavel**
Spojení	+420-549 491 810 +420-549 491 820 pary@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
42	Rychlý, Pavel - Kovář, Vojtěch. Displaying Bidirectional Text Concordances in KWIC format. In Proceedings of 5th Biennial Conference of the Asian Association for Lexicography. Chennai, India : University of Madras, 2007. s. 96-100.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/43/2007**

Název výsledku

Sborník konference First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007.

Abstrakt

The RASLAN Workshop is an event dedicated to exchange of information between research teams working on projects of computer processing of Slavonic languages and related areas. RASLAN is focused on theoretical as well as technical aspects of the project work, presentations of verified methods are welcomed together with descriptions of development trends. The workshop also serves as a place for discussion about new ideas. Topics of the Workshop will include (but are not limited to): * text corpora and tagging * syntactic parsing * sense disambiguation * machine translation * semantic networks and ontologies * semantic web * knowledge representation * applied systems and software This book constitutes the Proceedings of the First Workshop held in Karlova Studánka, Czech Republic, in December 2007. The 14 papers are organized in the proceedings targeted for researchers and advanced students in the areas of Morphological and Syntactic Parsing, Semantic Analysis, Text Processing Tools and Lexical Semantics.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr**
Spojení	+420-549 491 810 +420-549 491 820 sojka@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
43	Sojka, Petr - Horák, Aleš. First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007. Edited by Sojka P., Horák A. prvni. Brno : Masaryk University, 2007. 118 s. RASLAN Proceedings. ISBN 978-80-210-4471-5.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/44/2007**

Název výsledku

Multilingual Meta-Translator.

Abstrakt

This paper presents MetaTrans, a meta-search engine for online dictionaries. With this software, users are able to find translations in a number of online dictionaries simultaneously. The MetaTrans features a web interface which is easy to use. The modular design of the tool enables adding support for more online dictionaries with minimal effort. MetaTrans also utilizes information from text corpora, WordNets and a morphological analyzer.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pomikálek Jan**
Spojení	+420-549 491 810 +420-549 491 820 xpomikal@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
44	Pomikálek, Jan. MetaTrans -- Multilingual Meta-Translator. In RASLAN 2007 Proceedings. Brno : Masaryk University, Brno, 2007. od s. 109-115, 6 s. ISBN 978-1-56592-479-6	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/45/2007**

Název výsledku

Metodika provedení videozáznamu e-learningového kurzu a jeho využití.

Abstrakt

Častokrát se setkáváme s články nebo přednáškami, které pojednávají o pořizování záznamů z přednášek nebo jiných akcí. Obecně více odhalují chyby a nedostatky takto pořízených videí, než by podávaly návod, jak se těmto nedostatkům obecně vyhnout. Dokonce mnohdy je závěr takový. Že výsledná kvalita neodpovídá snaze autorů a vloženým prostředkům, takže je třeba v nejlepším přestat. Článek podává výsledky naší poměrně dlouhodobé práce a napovídá, jak se vyrovnat s problémy, které jsme již částečně nebo úplně odstranili.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr**
Spojení	+420-549 491 810 +420-549 491 820 sojka@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
45	Šiler, Pavel - Sojka, Petr. Metodika provedení videozáznamu e-learningového kurzu a jeho využití. In Sborník 4. ročníku konference o elektronické podpoře výuky SCO 2007. Brno : Masarykova univerzita, 2007. od s. 87-92, 6 s. ISBN 978-80-210-4296-4	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/46/2007**

Název výsledku

Platform for Full-Syntax Grammar Development Using Meta-grammar Constructs

Abstrakt

This paper describes a combination of tools necessary for full or deep syntactic parsing of natural language -- the syntactic parser synt, the graphical Grammar Development Workbench, GDW and the VerbaLex verb valency lexicon tools. We describe the development of the mentioned tools and how they integrate into one system that allows a team of experts (computational linguists as well as programmers) to cooperate on the development of grammar covering all Czech language phenomena.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš**
Spojení	+420-549 491 810 +420-549 491 820 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
46	Horák, Aleš - Kadlec, Vladimír. Platform for Full-Syntax Grammar Development Using Meta-grammar Constructs. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation. 2006. vyd. Beijing, China : Tsinghua University Press, 2006. od s. 311-318, 8 s. ISBN 7-302-14060-X	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/47/2007**

Název výsledku

Automatic Analysis and Evaluation of a Domain-Specific Text Corpus

Abstrakt

Automatic analysis of domain-specific dialogues is a special part of common analysis of natural language texts. In this paper, we describe the creation of fundamental resource for working with dialogues about electrical power networks - the corpus of 1 million tokens specialized to the power networks topics. We show the details of building such corpus and results of automatic analysis of the corpus content such as the term extraction, morphological disambiguation and syntactic analysis of the domain-specific texts.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš**
Spojení	+420-549 491 810 +420-549 491 820 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
47	Kovář, Vojtěch - Horák, Aleš. Power Networks Dialogues - Automatic Analysis and Evaluation of a Domain-Specific Text Corpus. In Proceedings of ELNET 2007. Ostrava : Faculty of Electrical Engineering and Computer Science, VŠB - Technical University of Ostrava, 2007. s. 30-37. ISBN 978-80-248-1681-4	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/48/2007**

Název výsledku

Sborník konference SCO 2007, Sharable Content Objects

Abstrakt

Kniha obsahuje 37 recenzovaných příspěvků konference, abstrakty dvou zvaných přednášek, autorský a tematický rejstřík. Články jsou ve sborníku členěny do osmi tematických oblastí. Kniha je určena všem zájemcům o aktuální přístupy a výzkum v oblasti elektronické podpory výuky -- e-learningu. Přílohou je CD-ROM s elektronickou hypertextovou podobou všech článků, elektronickou výstavkou kurzů a webem konference.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr**
Spojení	+420-549 491 810 +420-549 491 820 sojka@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
48	Sojka, Petr - Kvizda, Martin. SCO 2007, Sharable Content Objects, 4. ročník konference o elektronické podpoře výuky, Brno, Česká republika. Edited by Sojka P., Kvizda M. prvni. Brno : Masaryk University, 2007. 260 s. SCO Proceedings. ISBN 978-80-210-4296-4	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/49/2007**

Název výsledku

The Development of a Complex-Structured Lexicon based on WordNet

Abstrakt

The Cornetto project develops a new complex-structured lexicon for the Dutch language. The lexicon comprises information from two current electronic dictionaries - the Referentie Bestand Nederlands (RBN), which contains FrameNet-like structures, and the Dutch WordNet (DWN) with the usual WordNet structures. The Cornetto lexicon (stored in the Cornetto database) will be linked to English WordNet synsets and have detailed descriptions of lexical items in terms of morphologic, syntactic, combinatoric and semantic information. The database is organized in four data collections - lexical units, synsets, ontology terms and the Cornetto identifiers. The Cornetto identifiers are specifically used for managing the relations between lexical units on the one hand and synsets on the other hand. The mapping is first created automatically, but then revised manually by lexicographers. Special interfaces have been developed to compare the different perspectives of organizing concepts (lexical units versus synsets versus ontology terms). In this article, we describe the background information about the Cornetto project and the implementation of necessary project tools that are based on the DEBVisDic tool for WordNet editing. The development of the Cornetto clients is a joint project of the Masaryk University in Brno and the University of Amsterdam.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš**
Spojení	+420-549 491 810 +420-549 491 820 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
49	Horák, Aleš - Vossen, Piek - Rambousek, Adam. The Development of a Complex-Structured Lexicon based on WordNet. In Proceedings of the Fourth Global WordNet Conference. Szeged : University of Szeged, 2008. od s. 200-208, 9 s. ISBN 978-963-482-854-9	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/50/2007**

Název výsledku

The Global WordNet Grid Software Design

Abstrakt

In the presented paper we show how the Global WordNet Grid software is designed. The goal of Grid is to provide a free network of WordNets linked together through interlingual indexes. We have set as our goal to work on the Grid preparation in the Masaryk University NLP Centre and design its software background. All participating WordNets will be encapsulated by a DEB (Dictionary Editor and Browser) server established for this purpose. The following text presents design details of the new DEBGrid application with possibilities of three types of public and authenticated user access to the Grid WordNet data.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš**
Spojení	+420-549 491 810 +420-549 491 820 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita, Fakulta informatiky Botanická 68a 60200 Brno www.fi.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
50	Horák, Aleš - Pala, Karel - Rambousek, Adam. The Global WordNet Grid Software Design. In Proceedings of the Fourth Global WordNet Conference. Szeged : University of Szeged, 2008. od s. 194-199, 6 s. ISBN 978-963-482-854-9.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

4.1.3. PLNĚNÍ DÍLČÍCH CÍLŮ

4.1.3.1. ZPRÁVA O DOSAŽENÍ DÍLČÍHO CÍLE

Číslo dílčího cíle	1
Název dílčího cíle	Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím.
Plánované datum dosažení dílčího cíle	31.12.2007

INDIKÁTORY DOSAŽENÍ VÝSTUPU - SKUTEČNĚ DOSAŽENÉ

V průběhu roku 2007 byly vytvořeny následující řečové korpusy, které jsou využívány především pro trénování recognizeru spontánní české řeči:

1) LAC-SS - LICS Audio Corpus/Spontaneous Speech
Korpus LAC-SS je korpus nahrávek spontánní řeči opatřený rozšířenou synchronní ortografickou transkripcí. V současné době obsahuje LAC-SS celkem 741 minut (12h21m45s) záznamu spontánní řeči ve 24 nahrávkách vysoké kvality, k nimž byla pořízena manuální transkripce. 11 nahrávek pochází z přednášek pracovníků Katedry informatiky a výpočetní techniky - jedná se tedy o mluvčí s rozsáhlou praxí v oblasti veřejných projevů. Naopak 13 nahrávek pochází ze studentských seminářů - mluvčí (studenti) ve většině případů nejsou zvyklí pronášet veřejné projevy, a proto také záznamy obsahují celou řadu neřečových zvuků a projevů např. nervozity, což je v tomto případě ale žádoucí, protože díky tomu je korpus dostatečně foneticky bohatý.
Z pohledu výpočetní lingvistiky korpus obsahuje celkem 40866 lexikálních atomů, z toho 33795 slov a 7071 neřečových zvuků (non-speech sounds, např. odkašlávání, popotahování, kýchání, mlaskání, apod.). Celkový počet různých slov v korpusu je 6894.
Podrobnější statistika získaného materiálu je k dispozici v článku [Ekštein K.: On Building of Czech Spontaneous Speech Corpus, SPECOM 2007]. Korpus je k dispozici na adrese http://liks.fav.zcu.cz/mediawiki/index.php/LAC-SS .

2) Sémantický korpus dotazů
Vytvořený korpus obsahuje 20292 tématicky zaměřených dotazů uživatelů určených pro internetový vyhledávač. Všechny dotazy jsou zpracovány týmem vyškolených anotačních pracovníků. Celá množina dotazů byla tématicky roztříděna a neužitečné dotazy byly odfiltrovány. Mezi anotatátorská shoda pro určování témat dosahovala 83%.
Podmnožina 6750 vět byla sémanticky anotována, tj. byl u nich určen a zaznamenán význam dle specifikované metodologie. Mezianotátorská shoda dosáhla pro tyto dotazy hodnotu 75%. Všechny dotazy byly vždy anotovány dvěma pracovníky a sporné výsledky následně zjednoznačněny koordinátorem anotačních prací. Korpus ke k dispozici na adrese http://liks.fav.zcu.cz/mediawiki/index.php/SemKorp .

3) Programový systém CzechWebSOM pro kategorizaci česky psaných dokumentů
Implementace systému WEBSOM - kolekce metod pro zpracování sémantiky a kategorizaci kolekce českých dokumentů - v jazyce Java přináší řadu výhod vyplývajících z objektově orientovaného návrhu, vyšší úrovně abstrakce jazyka a možnosti integrace softwarů třetích stran. Software umožňuje vytvoření číselné reprezentace kolekce dokumentů, redukci dimenze dat a zpracování dat dvouvrstvou architekturou Kohonenovy mapy (mapa slovních kategorií, mapa dokumentů). K dispozici je i grafické uživatelské rozhraní, které umožňuje uživatelsky jednoduché a příjemné nastavení parametrů metody a provedení většího množství experimentů. Systém je k dispozici na adrese http://liks.fav.zcu.cz/mediawiki/index.php/WEBSOM .

4)Kolekce dat pro dolování ze struktury Webu
Korpus je přístupný přes http://textmining.zcu.cz/
Obsahuje korpus dokumentů stažených z českých kateder informatiky. PDF a PS soubory byly převedeny do textu.

Dále obsahuje korpus dokumentů stažených z francouzských kateder informatiky. PDF a PS soubory byly převedeny do textu. V adresáři Fr-Statistics jsou statiskiky o běhu programu pro skupiny o čtyřech katedrách a příslušné databáze SQLite.
Korpus dokumentu stažených ze slovenských kateder informatiky. PDF a PS soubory byly převedeny do textu.

5)Korpus závadných textů
Korpus není přístupný vzhledem k právním zábranám a slouží jen pro testovací účely. Aby mohlo být dosaženo dobrých výsledků při filtraci dat realizované textovými klasifikátory, je potřeba mít při jejich trénování k dispozici nejen dokumenty patřící do rozpoznávané množiny, ale i dokumenty patřící do množiny opačné. Proto vznikl za pomoci studentů a nástrojů vytvořených v rámci aktivit 17/2006 a 19/2006 vícejazyčný dataset závadných textových dokumentů pokrývající oblast pornografie, a to v českém, slovenském, francouzském a německém jazyce. V souladu s výše zmíněnými potřebami klasifikátorů textu je ke každému jazyku vždy k dispozici jak množina závadných, tak i nezávadných dat. Dataset obsahuje pro každý jazyk 200 dokumentů, z toho 100 dokumentů pokrývá závadnou a 100 nezávadnou oblast. Data jsme se rozhodli ponechat v původním HTML formátu. Díky tomu je možné zohlednit váhu textu uvozeného různými HTML tagy a zlepšit tak případně poměr mezi správně a chybně filtrovanými dokumenty. K dispozici je i seznam internetových serverů, ze kterých byly jednotlivé dokumenty získány, což umožňuje v případě potřeby vytvořit rozsáhlejší datový korpus pokrývající zmíněnou oblast.
6)Korpusy pro vícejazykové zpracování
Jsou přístupné přes http://textmining.zcu.cz/
Korpus Reuters a ČTK předzpracovaný EWN Vytvoření korpusu vhodného pro testování prototypových řešení Velikost: 480MB, 5 klasifikačních tříd po 8000 dokumentech Popis: Vytvořený korpus obsahuje texty tiskových agentur Reuters a ČTK předzpracované různými metodami lemmatizace a indexace. Jedna z metod je indexace pomocí EWN. Korpus je speciálně určen pro testování multilinguálních metod.
Korpus Die Welt
Velikost: 32MB, 1000 dokumentů Popis: Korpus vznikl ověřením správné funkce při řešení aktivity "Návrh systému pro tvorbu korpusů z webu". Byl vygenerovaný automaticky podle uživatelsky zadaných pravidel. Vícejazykový korpus tiskových zpráv uložených na webu Je výsledkem doplnění jazykových korpusů o další evropské jazyky Velikost: 500MB data, vybraných 20MB pro ruční extrakci užitečné informace
Popis: Vícejazykový korpus obsahující data z významných českých, německých a anglických informačních serverů. Část korpusu označená "Deutsche Welle - Multilingual" obsahuje články ve většině významných evropských jazycích a vybraných asijských. Součástí korpusu jsou šablony použité pro extrakci užitečné informace a ručně vytvořené extrakty
dokumentů. Korpus je vhodný pro ověření algoritmů racujících s vícejazyčnými daty.

7)Kolekce českých textů pro multidokumentovou sumarizaci
Korpus je přístupný přes http://textmining.zcu.cz/
Z důvodu neexistence kolekce českých dokumentů anotovaných pro sumarizaci byl vytvořen sumarizační korpus. Tento korpus obsahuje 7 shluků dokumentů (80k slov). Každý shluk se týká určité události/tématu. Navíc každý shluk obsahuje 3 dotazy (úlohy) – jeden obecný, jeden specifický a jeden inkrementální. 4 anotátoři vytvořili jak abstrakty (souhrnem je text napsaný anotátorem), tak extrakty (souhrnem jsou vybrané věty původních dokumentů) pro každý shluk a dotaz. Vše je ukládáno v XML formátu. Korpus byl použit pro testování Multi-document sumarizátoru založeném na LSA Na základě zaslání požadavku ze stránek textmining.zcu.cz bude korpus poskytnut ke stažení.

8)Korpus syntaktických stromů
Pro účely automatického statistického vyhodnocení kvality vyvíjených algoritmů pro uspořádání výsledků syntaktické analýzy je zapotřebí připravit co nejrozsáhlejší korpus syntaktických frázových stromů systému synt. V roce 2007 jsme připravili základní korpus obsahující cca 5000 syntaktických stromů, který plánujeme dále rozšiřovat. Vstupní věty korpusu jsou založeny na pražském korpusu PDT-1.0, aby bylo možné syntaktickou informaci srovnávat s výsledky pražské školy, jejíž algoritmy i výstupy jsou založeny na jiných principech a

algoritmech (složkové vs. závislostní stromy, stochastické vs. pravidlové systémy).
Korpus je dostupný v textovém formátu vhodném pro zpracování externími nástroji, pomocí korpusového manažeru pro vyhledávání a pomocí editoru syntaktických stromů pro další anotaci.

9)Korpus vzorových přepisů vybraných vět a jejich sémantické reprezentace
Pro účely logické analýzy vět v přirozeném jazyce jsme vytvořili testovací sadu vzorových (typových) přepisů českých vět do konstrukcí transparentní intenzionální logiky (TIL) spolu s přiřazením odpovídajících sémantických reprezentací. Uvedená sada obsahuje v současnosti jen základní kolekci vět pro testování inference v systému Dolphin. Tento systém je první prototyp implementující bázi znalostí TIL. Pro tvorbu konstrukcí se využívá valenční lexikon VerbaLex, který je také rozšiřován v rámci projektu.
Vytvořený korpus je přístupný jako sada konstrukcí TILu ve formátu pro vstup systému Dolphin, což je zároveň formát, který poskytuje systém synt v případě automatické tvorby logických konstrukcí.
10)Lexikální databáze Verbalex obsahující valenční rámce českých sloves
Verze databáze dokončená v roce 2007 čítá přes 10000 českých sloves a 28000 rámců. Rámce jsou napojeny na inventář asi 240 ontologických kategorií, sloužících jako sémantické role.
Databáze je dostupná na adrese http://nlp.fi.muni.cz/verbalex/

PROSTŘEDKY OVĚŘENÍ VÝSTUPU - SKUTEČNĚ DOSAŽENÉ

Korpusy budou v následujícím období nadále doplňovány a používány k pokusnému trénování vyvíjeného systému pro rozpoznávání přirozeně pronesených promluv z různých tématických oblastí. K současnému datu bylo ověřeno, že v závislosti na dialogové doméně dosahuje vyvinutý recognizer po natrénování uvedenými korpusy spolehlivost rozpoznávání jednotlivých slov 83 - 97 %, spolehlivost rozpoznávání a porozumění obsahu celých vět se pohybuje zhruba okolo 95 %.

4.1.4. REDAKČNĚ UPRAVENÁ ZPRÁVA

Projekt má za cíl vývoj nástrojů pro komunikaci s Webem v přirozeném jazyce. V r. 2007 byly vytvářeny korpusy pro testování navrhovaných algoritmů, implementovány výchozí a podpůrné algoritmy pro zpracování obsahu Webu prostřednictvím přirozeného jazyka.

4.1.5. PLNĚNÍ PODMÍNEK PROGRAMU

Plnění specifických podmínek programu - se pro projekty NPV II nezpracovává. Pro projekty NPVII specifické podmínky ve vyhlášení programu nebyly formulovány.

4.1.6. PLNĚNÍ SMLOUVY O SPOLUPRÁCI

Na základě vymezených základních práv (viz uzavřená smlouva upravující vztahy mezi příjemcem a spolupříjemcem) příjemce poskytnul spolupříjemci finanční dotaci přímým převodem na stanovený účet Masarykovy univerzity, náklady na projekt byly vedeny v oddělené evidenci obou spolupracujících subjektů.
Uzavřená smlouva o spolupráci je plněna beze zbytku, plánované finanční prostředky byly vyčerpány - viz odstavec 2.3.2.

4.2. DALŠÍ PŘÍLOHY - rok 2007

4.2.1. Odborné a věcné přílohy zprávy - seznam

	Pořadí	Soubor
	1	**Seznam publikací a softwaru - Plzeň** Seznam publikací, které se v roce 2007 dotýkaly řešeného projektu, a dále seznam zpracovaných korpusů a vyvinutého softwaru (vše za rok 2007) Publikace_Plzen.doc *(62 kB)*
	2	**Seznam publikací - Brno** Seznam publikací za rok 2007 - pracoviště MU Brno Publikace_Brno.doc *(27 kB)*
	3	**Horák, Aleš - Pala, Karel. Building a Large Lexicon of Complex Valency Frames. In Proceedings of the FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages. Tartu, Estonia : Lund University, Sweden, 2007.** Článek o budování rozsáhlého lexikonu komplexních valenčních rámců VerbaLex frame2007_pala_hales.pdf *(87 kB)*
	4	**Radek Vykydal, Grammar Development Workbench - manuál** Systém GDW je prostředí pro vývoj gramatik přirozeného jazyka postavené nad systémem synt vyvinutým v Laboratoři NLP na Fakultě Informatiky Masarykovy univerzity v Brně. Jeho autorem, stejně jako autorem tohoto manuálu, je Radek Vykydal, který na projektu pracoval pod vedením Aleše Horáka. Systém byl vyvinut a použit mimo jiné pro tvorbu korpusu složkových stromů systému synt. Celou dokumentaci nebylo možné do systému nahrát, přikládáme tedy pouze popis modulu TreeView. gdw_male.pdf *(4862 kB)*
	5	**Andrej Gardoň, Návrh báze znalostí a základního inferenčního stroje pro TIL, bakalářská práce** Text obsahuje podrobný popis první prototypové implementace systému Dolphin s využitím testovací sady vzorových (typových) přepisů českých vět do konstrukcí transparentní intenzionální logiky (TIL) včetně přiřazením odpovídajících sémantických reprezentací. dolphin.pdf *(447 kB)*

4.2.2. Ostatní (např. možné využití výsledků) - seznam

	Pořadí	Soubor
	1	*(kB)*

4.2.3. Zápisy z projednání (oponentské posudky) - seznam

	Pořadí	Soubor
		V elektronické podobě soubor nebyl řešitelským týmem poskytnut.

4.2.4. Zápisy a dokumenty z jednání se styčnými pracovníky zadavatele - seznam

	Pořadí	Soubor
		V elektronické podobě soubor nebyl řešitelským týmem poskytnut.

4.2.5. Zápisy z jednání Rady projektu (Centra) - seznam

Příloha 4.2.5. Zápisy z jednání Rady projektu (Centra) - se pro tento program nezpracovává.

4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodněním - seznam

Příloha 4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodnění - se pro tento program nezpracovává.